Summer 6-21-2021

Creation and Application of Various Tools for the Reconstruction, Curation, and Analysis of Genome-Scale Models of Metabolism

Wheaton L. Schroeder
University of Nebraska - Lincoln, wheaton@huskers.unl.edu

CREATION AND APPLICATION OF VARIOUS TOOLS FOR THE RECONSTRUCTION, CURATION, AND ANALYSIS OF GENOME-SCALE MODELS OF METABOLISM

by

Wheaton L. Schroeder

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Chemical and Biomolecular Engineering

Under the Supervision of Professor Rajib Saha

Lincoln, Nebraska

June 21, 2021

CREATION AND APPLICATION OF VARIOUS TOOLS FOR THE RECONSTRUCTION, CURATION, AND ANALYSIS OF GENOME-SCALE MODELS OF METABOLISM

Wheaton L. Schroeder, Ph.D.

University of Nebraska, 2021

Advisor: Rajib Saha

Systems biology uses mathematics tools, modeling, and analysis for holistic understanding and design of biological systems, allowing the investigation of metabolism and the generation of actionable hypotheses based on model analyses. Detailed here are several systems biology tools for model reconstruction, curation, analysis, and application through synthetic biology. The first, OptFill, is a holistic (whole model) and conservative (minimizing change) tool to aid in genome-scale model (GSM) reconstructions by filling metabolic gaps caused by lack of system knowledge. This is accomplished through Mixed Integer Linear Programming (MILP), one step of which may also be independently used as an additional curation tool. OptFill is applied to a GSM reconstruction of the melanized fungus *Exophiala dermatitidis*, which underwent various analyses investigating pigmentogenesis and similarity to human melanogenesis. Analysis suggest that carotenoids serve a currently unknown function in *E. dermatitidis* and that *E. dermatitidis* could serve as a model of human melanocytes for biomedical applications. Next, a new approach to dynamic Flux Balance Analysis (dFBA) is detailed, the Optimization- and Runge-Kutta- based Approach (ORKA). The ORKA is applied to the model plant *Arabidopsis thaliana* to show its ability to recreate *in vivo* observations. The analyzed model is more detailed than previous models, encompassing a larger time scale, modeling more tissues, and with higher accuracy. Finally, a pair of tools, the Eukaryotic Genetic Circuit Design (EuGeneCiD) and Modeling (EuGeneCiM) tools, is introduced which can aid in the design and modeling of synthetic biology applications

hypothesized using systems biology. These tools bring a computational approach to synthetic biology, and are applied to *Arabidopsis thaliana* to design thousands of potential two-input genetic circuits which satisfy 27 different input and logic gate combinations. EuGeneCiM is further used to model a repressilator circuit. Efforts are ongoing to disseminate these tools to maximize their impact on the field of systems biology. Future research will include further investigation of *E. dermatitidis* through modeling and expanding my expertise to kinetic models of metabolism.

ACKNOWLEDGEMENTS

This dissertation would not have been possible without two people. The first is my advisor, Dr. Rajib Saha, who took me as a graduate student in 2017 when I was let go by my previous advisor due to their lack of funding. Without his offer of a second chance, I may never have published articles or completed this dissertation. Throughout my time as a graduate student, he as encouraged me to find my own career path and has provided sage advice. Whatever my successes be after graduation, they will owe much to him. For these reasons he has my heartfelt gratitude. The second person is my wife, Christine, without whose constant support the stress, difficulties, and uncertainties inherent in the process of seeking a Doctorate of Philosophy would have been much more difficult to bare. I would also like to take this opportunity to acknowledge her sacrifices in getting me to this point, including allowing my education and future career to dictate where we live.

I would like to acknowledge the graduate student and postdoctoral members of the Systems and Synthetic Biology (SSBio) laboratory for aiding me in many ways including my mental health (through social activities), my activities as a graduate student (though reviewing my papers and helping generate/refine research ideas), my career (sharing job postings, career advice, and career development opportunities). Specifically, these are Dr. Cheryl Immethun (postdoctoral scholar), Mohammad Mazharul Islam (graduate student), Adil Al-Siyabi (graduate student), Dianna Long (graduate student), Brandi Brown (graduate student), Mark Kathol (graduate student), and Niaz Chowdhury (graduate student).

I would also like to acknowledge the undergraduate and high school researchers who have contributed to my work. This includes Joshua C. Bauman (UNL undergraduate), Ali Keshk

(Lincoln Southwest High School), Bree Brunsman (Lincoln East High School), Kareem Keshk (Lincoln Southwest High School), and Conner William Lunn (UNL undergraduate).

In addition, I would like to acknowledge the Holland Computing Center for providing the supercomputing cluster and resources necessary to conduct my research, and for providing technical support when needed.

I would finally like to acknowledge the funding organizations which have supported my work. The first is the University of Nebraska – Lincoln (UNL) for the generous support of my work through my advisor's Faculty Startup Grant (21-1106-4038). Later in my time as a graduate student, I was generously supported by a grant from the NSF EPSCoR Center for Root and Rhizobiome Innovation (CRRI, 25-1215-0139-025).

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

1. BACKGROUND, CONTEXT, AND DISSERTATION GOALS

Portions of this material have previously appeared in the following publication:

W. L. Schroeder, R. Saha, OptFill: A Tool for Infeasible Cycle-Free Gapfilling of Stoichiometric Metabolic Models, iScience, 23(2020) 1-14. Used with permission.

W. L. Schroeder, S. D. Harris, and R. Saha, Computation-Driven Analysis of Model Polyextremotolerant Fungus Exophiala dermatitidis: Defensive Pigment Metabolic Costs and Human Applications, iScience, 23(2020) 1-17. Used with permission.

W. L. Schroeder, R. Saha, Introducing an Optimization- and explicit Runge-Kutta- based Approach to Perform Dynamic Flux Balance Analysis, Scientific Reports, 10:9241(2020) 1-28. Used with permission.

W. L. Schroeder, R. Saha, Protocol for Genome-Scale Reconstruction and Melanogenesis Analysis of Exophiala dermatitidis, STAR Protocols, 1(2020) 1-37. Used with permission.

1.1. PREFACE

This chapter is designed to provide three sets of knowledge to the reader which may be necessary to the understanding and critical analysis of this dissertation, particularly for non-subject matter experts. The first section (background) will be to provide the reader with background knowledge related to the field of Systems Biology and specifically to the various types of Systems Biology concepts, tools, and terminologies used throughout this dissertation. The second section (context) will provide context for the works appearing in this dissertation so that the novelty of presented work may be evident. The third section (dissertation goals) will provide an overview of the goals of the dissertation research and attempt to unify these chapters under a common framework.

1.2. BACKGROUND

The field of systems biology, which is the discipline central to this dissertation, is closely linked with that of synthetic biology which will be introduced first. Synthetic biology is the redesigning of organisms to accomplish specific tasks, often by the manipulation of an organism's genome. The use of synthetic biology for the engineering of uni- and multi-cellular organisms to enhance desirable phenotypes in microbe, plant, and animal systems, has been well established and has been capable of affecting the lives of millions of individuals, such as in the case of artemisinin production in yeast or enhancing nutritional value of agricultural products (Beyer et al., 2002; Hall et al., 2008). Synthetic biology techniques have been applied to many plant systems such as tomatoes (Gonzali, Mazzucato, & Perata, 2009), rice (Beyer et al., 2002), and maize (Gonzali et al., 2009) to produce enhanced phenotypes often with application to human nutrition (Hall et al., 2008), pest resistance (Hilder & Boulter, 1999), and resilience to abiotic stresses (T. H. H. Chen & Murata, 2002). Many of these efforts have focused on a genetic understanding and manipulation of the plant system (or plant tissue) in question, having relied on intuitive interventions such as changes in regulation, insertion of new gene(s), and deletion of gene(s) from competing pathway(s) (Hall et al., 2008; Hilder & Boulter, 1999; T. H. H. Chen & Murata, 2002).

Systems biology is the use of computational and mathematics tools, modeling, and analysis for holistic understanding and design of biological systems. Therefore, systems biology may be seen as a method of generating hypotheses *in silico* utilizing mathematics and knowledge of the biological system which may be investigated *in vivo* through synthetic biology or other more traditional methods. While systems biology has many possible applications and aspects, in this dissertation the focus will be on computational modeling of metabolism and various tool for building or analyzing these models. Here, metabolism is defined as the set of chemical reactions and exchanges which occur in a living system, whether that system be a single cell, organism, or group of organisms.

The most basic and commonly used form of systems biology model is the Stoichiometric Model (abbreviated as SM; a list of all abbreviations used can be found in the "acronyms used" section in chapter 6, though all abbreviations are still defined at first use) of metabolism, which has provided a more rigorous method of metabolic investigation (Thiele & Palsson, 2010; Orth et al., 2010). A SM is, essentially, a matrix of reaction stoichiometries representing the metabolism of an organism utilizing the stoichiometry of the chemical exchanges which is often represented using linear algebra. Shown below is the basic form for a stoichiometric model.

$$\begin{bmatrix} \boldsymbol{S_{11}} & \cdots & \boldsymbol{S_{1m}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{S_{m1}} & \cdots & \boldsymbol{S_{nm}} \end{bmatrix} \tag{1.1}$$

Where S_{ij} is the stoichiometric coefficient of metabolite i in reaction j. Note that it is the convention to represent the set of metabolites as I with elements i and represent the set of reactions as J with elements j. Note that this is inevitably a sparce matrix, non-square (e.g. $n \neq m$) matrix. The set of reactions in an SM is defined (in part) by the Gene-Protein-Reaction (GPR) links (Thiele & Palsson, 2010; Terzer et al., 2009) in an organism. To elaborate, when reconstructing such a model, publicly available databases (such as NCBI, KEGG, ModelSeed, and KBase among others) are used to identify which proteins are produced by an organism. These proteins are then investigated to determine what chemical reactions the organism can metabolize. When considering a large group of genes, proteins, and reaction the model is said to model metabolism (Thiele & Palsson, 2010; Terzer et al., 2009). Other reaction added to the SM include chemical exchanges across system boundaries (such as organelles, the cell membrane, and the cell external environment) and a pseudoreaction which is called the biomass equation which allows the model to simulate growth. Details on how this equation is formulation are given in Thiele & Palsson, 2010 and will not be expounded upon here since this is not central to the understanding of this dissertation. These GPR links allow for investigation of genetic effects on metabolism. When an SM encompasses the entire chemical reaction repertoire of an organism (as determined through genomic knowledge), it is called a Genome Scale Models (abbreviated as GSM or GEM, this work will use the former for preference) (Oberhardt, Palsson, & Papin, 2009)(Thiele & Palsson, 2010)

GSMs, since they are a matrix, do not operate independent of mathematical analysis methods. These analysis methods are almost universally optimization-based because, as the stoichiometric matrix is non-square, resulting in systems of equations which have differing numbers of variables and equations. Because of this any analysis performed on a GSM has a solution space, that is a range of values in n-dimensions which the variables of the system might take to solve the system of equations. Optimization is a method of choosing a "best" solution of this system of equations subject to a set of criteria and an objective function. To illustrate, Perhaps the most common analysis tool used with these models is Flux Balance Analysis (FBA) (Orth et al., 2010; Terzer et al., 2009). This tool seeks to calculate the rate of chemical reactions and exchanges occurring throughout the modeled organism, and essentially is a set of Ordinary Differential Equations (ODEs, see below).

$$\boldsymbol{Maximize\ v_{biomass}} \tag{1.2}$$

$$\boldsymbol{subject\ to\ (s.t.)}$$

$$\begin{bmatrix} \boldsymbol{S_{11}} & \cdots & \boldsymbol{S_{1m}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{S_{n1}} & \cdots & \boldsymbol{S_{nm}} \end{bmatrix} \begin{bmatrix} \boldsymbol{v_1} \\ \vdots \\ \boldsymbol{v_m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{dC_1/dt} \\ \vdots \\ \boldsymbol{dC_m/dt} \end{bmatrix} \tag{1.3}$$

Where v_j is the rate of reaction j or the metabolic flux through reaction j (both terminologies are used in this dissertation) and C_i is the concentration of metabolite i. In FBA, and throughout this dissertation, metabolic flux will have units of $mmol/gDW \cdot h$ where gDW is the dry mass of the organism. This unit normalizes flux units. The exception in $v_{biomass}$, the flux through the biomass pseudoreaction, which, because of its formulation, has units of h^{-1} and represents the doubling rate of the organism. Note that shown in equation (1.2) is an objective to maximize organism growth by maximizing the reaction flux through the biomass pseudoreaction. Other growth objectives are possible and commonly used including minimizing uptake of some nutrient (de Oliveira Dal'Molin et al., 2010; Gomes de Oliveira Dal'Molin et al., 2015), maximizing growth (Orth et al., 2010), or maximizing a desired bioproduct (Terzer et al., 2009). The formulation show above is the dynamic Flux Balance Analysis (dFBA), since it allows for changes in metabolism with respect to time (this is the focus of chapter 4). The basic FBA tool simplifies equation (1.3) by applying the Pseudosteady State Hypothesis (PSSH), which assumes that the system does not change with time. The FBA formulation is shown below.

$$\boldsymbol{Maximize\ v_{biomass}} \tag{1.4}$$

$$\boldsymbol{s.t.}$$

$$\begin{bmatrix} \boldsymbol{S_{11}} & \cdots & \boldsymbol{S_{1m}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{S_{n1}} & \cdots & \boldsymbol{S_{nm}} \end{bmatrix} \begin{bmatrix} \boldsymbol{v_1} \\ \vdots \\ \boldsymbol{v_m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} \\ \vdots \\ \boldsymbol{0} \end{bmatrix} \tag{1.5}$$

$$\boldsymbol{Other\ constraints}$$

This simplifies the analysis to a Linear Programming (LP) problem. FBA can find the extremum of a given growth objective which is defined by an objective function subject to mass balance, reaction directionality, and certain other constraints (generally restricting growth rate or nutrient uptake depending upon the objective function) (Thiele & Palsson, 2010; Orth et al., 2010;

Terzer et al., 2009; Gomes de Oliveira Dal'Molin et al., 2015). Note that other constraints are generally added to the optimization problem such as limiting the update of essential nutrients, biomass production, or other constraints.

Another ubiquitous tool applied to GSMs is Flux Variability Analysis (FVA). The basic formulation is the same as FBA; however, the formulation is solved for each reaction and for both maximal and minimal values.

$$\boldsymbol{foreach\, j \in J}:$$

$$\boldsymbol{Maximize\ v_j} \tag{1.6}$$

$$\boldsymbol{s.t.} \tag{1.7}$$

$$\begin{bmatrix} \boldsymbol{S_{11}} & \cdots & \boldsymbol{S_{1m}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{S_{n1}} & \cdots & \boldsymbol{S_{nm}} \end{bmatrix} \begin{bmatrix} \boldsymbol{v_1} \\ \vdots \\ \boldsymbol{v_m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} \\ \vdots \\ \boldsymbol{0} \end{bmatrix} \tag{1.8}$$

$$\boldsymbol{Other\ constraints}$$

and

$$\boldsymbol{Minimize\ v_j} \tag{1.9}$$

$$\boldsymbol{s.t.} \tag{1.10}$$

$$\begin{bmatrix} \boldsymbol{S_{11}} & \cdots & \boldsymbol{S_{1m}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{S_{n1}} & \cdots & \boldsymbol{S_{nm}} \end{bmatrix} \begin{bmatrix} \boldsymbol{v_1} \\ \vdots \\ \boldsymbol{v_m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} \\ \vdots \\ \boldsymbol{0} \end{bmatrix} \tag{1.11}$$

$$\boldsymbol{Other\ constraints}$$

FVA therefore identifies the ranges of values which each variable (v_j) can take in the set of equations defined by FBA. This can be useful in identifying various issues which require model curation or the definition of additional constraints.

GSMs, with their associated analysis tools such as FBA and FVA, have become an indispensable tool of systems biology in a wide variety of applications (Thiele & Palsson, 2010), with perhaps the most common applications being the overproduction of a native metabolite (Khodayari, Chowdhury and Maranas, 2015; Lin *et al.*, 2017; Zhang, Tervo and Reed, 2016; Feist and Palsson, 2008) or engineering of metabolism to produce a non-native metabolite (Feist and Palsson, 2008; Gudmundsson, Agudo and Nogales, 2017a; Gudmundsson, Agudo and Nogales, 2017b). Other uses of GSMs have also been to characterize Open Reading Frames (ORFs), determine gene essentiality, and evolutionary studies in *Escherichia coli* (Feist & Palsson, 2008); investigate the Warburg effect and drug screenings in human cancer cells (Yizhak, Chaneton, Gottlieb, & Ruppin, 2015); study interactions among members of a microbial community (Stolyar *et al.*, 2007; Magnúsdóttir *et al.*, 2016); and to investigate plant metabolism under stress (Cheung *et al.*, 2013; Williams *et al.*, 2010; Cramer *et al.*, 2011).

1.3. CONTEXT

Some of the first SMs of *Escherichia coli* were published in 1990, with the first true GSMs only possible after the *E. coli* genome was sequenced in 1997. (Reed & Palsson, 2003) By the turn of the millennium, only a handful of GSMs had been reconstructed, and by 2008 there were 45 GSM reconstructions representing more than 30 species. (T. Y. Kim, Sohn, Kim, Kim, & Lee, 2012) It was around this time, in 2010, that the now standard protocol was published for GMS reconstructions. (Thiele & Palsson, 2010) As the number of reconstructions, and species reconstructed has grown exponentially, resulting, as of February 2019, in 6239 organism with at least one GSM (Gu, Kim, Kim, Kim, & Lee, 2019).

However, there are limitations to this breadth and diversity. For instance, the vast majority, approximately 94%, of GSMs (as of February 2019) identified by Gu et al. (2019) were of prokaryotic organisms. This is for several reasons: 1) prokaryotes have simpler and smaller genomes (important for, sequencing and annotating genomes as well as reconstructing GSMs), 2) prokaryotes are able to make use of novel feedstocks or produce novel and valuable bioproducts, 3) multicellular organisms are more complex to model, and 4) prokaryotes are abundant. Setting aside these advantages for the ease of modeling, eukaryotic models have important implications for human health and for understanding multiscale metabolism, among many other applications, and should not be ignored. Chapters 3 and 4 highlight the creation of two eukaryotic metabolic models (the former is a GSM, the latter an SM) and a new mathematical approach to studying dynamic metabolism.

According to the standard protocol, GSMs are very time and labor intensive to produce, taking between six months to several years of manpower to reconstruct (Thiele & Palsson, 2010). One particularly important, and time-consuming, obstacle to GSM reconstruction is identifiable, though not easily addressed, utilizing FVA which is the problem of Thermodynamically Infeasible Cycles (TICs) which might also be called futile cycles or type III cycling. TICs result from the mathematics of FBA and FVA in that all reaction rates and linearly related and there are no kinetics-based limitations which impose limits on reaction rate (e.g. enzyme concentration or substrate concentration). Therefore, if two or more reactions sum to zero (such as $1A \rightarrow 1B$ and $1 \rightarrow 1A$), then this set of reactions can hold any flux value. This is identified by FVA when flux value of infeasibly large magnitude are identified. These become problematic to identify and resolve when more than one TIC exists, especially TICs consist of 4 or more reactions. Chapter 2 will address this issue in detail and show an optimization-based tool to address this issue.

Various automated tools for GSM reconstruction such as KBase (Arkin et al., 2018) and ModelSeed (Henry et al., 2010) have been developed to address this issue, and can effectively generate draft models as a starting point for GSM reconstructions. The draft models often need to be used cautiously and carefully curated however, since the often have several issues such as: 1) chemical or charge imbalance in reaction stoichiometries, 2) many reactions are often disconnected from the metabolic network, 3) some reactions are included in models with little to no evidence, 4) draft models often contain TICs, and 5) draft models are often overly generic and missing metabolic functions unique to an organism, family, or other taxonomic group. In fact, these automated tools do little more than reduce GSM reconstruction time a few days or weeks.

1.4. DISSERTATION GOALS

This dissertation covers several aspects of genome-scale modelling, including model reconstruction, curation, and analysis. The common theme of most the research presented here is the creation of tools for the curation (akin to "debugging") of GSM (OptFill and its component TIC-Finding Problem, Chapter 2) and analysis (the Optimization- and Runge-Kutta- based Approach, or ORKA, to dynamic FBA, Chapter 4). The goal of the curation tools is to provide a computational and rigorous resource which will increase the speed of model generation, ideally allowing the more frequent application of genome-scale modeling to more eukaryotic species and systems (such as communities). This is demonstrated in the reconstruction of a poorly studied, yet potentially important, melanized fungus, *Exophiala dermatitidis* (Chapter 3). The goal of the improved analysis tool (ORKA) is to provide a more numerically accurate and stable method by which to analyze dynamic and multi-tissue metabolism, which is more crucial for the study of eukaryotic systems. This tool is both introduced and applied, to the model plant *Arabidopsis thaliana*, to demonstrate its usefulness (Chapter 4).

To design and model synthetic biology applications based on the results of genome-scale modeling, an additional pair of computational tools which I have developed will be discussed in Chapter 5. These tools are the Eukaryotic Genetic Circuit Design (EuGeneCiD) and Modeling (EuGeneCiM) tools. These tools do not directly utilize genome-scale modeling, yet were designed using familiar computational methods, namely optimization and Mixed Integer Linear Programming, to generate and simulate potential genetic circuit designs from a provided database of bioparts (promotors, genes, transcripts, terminators, and enzyme). EuGeneCiD and EuGeneCiM will increase the speed of the design, modeling, and screening of future synthetic biology applications, as well as provide the benefits of optimization-based approaches such as the ability to identify best circuits and effective yet non-intuitive designs.

Chapter 2

2. OPTFILL: A TOOL FOR INFEASIBLE CYCLE-FREE GAPFILLING OF STOICHIOMETRIC METABOLIC MODELS

Portions of this material have previously appeared in the following publication:

W. L. Schroeder, R. Saha, OptFill: A Tool for Infeasible Cycle-Free Gapfilling of Stoichiometric Metabolic Models, iScience, 23(2020) 1-14. Used with permission.

W. L. Schroeder, S. D. Harris, and R. Saha, Computation-Driven Analysis of Model Polyextremotolerant Fungus Exophiala dermatitidis: Defensive Pigment Metabolic Costs and Human Applications, iScience, 23(2020) 1-17. Used with permission.

W. L. Schroeder, R. Saha, Protocol for Genome-Scale Reconstruction and Melanogenesis Analysis of Exophiala dermatitidis, STAR Protocols, 1(2020) 1-37. Used with permission.

2.1. PREFACE

Stoichiometric metabolic modeling, particularly genome-scale models (GSMs), is now an indispensable tool for systems biology. The model reconstruction process typically involves collecting information from public databases; however, incomplete systems knowledge leaves gaps in any reconstruction. Current tools for addressing gaps use databases of biochemical functionalities to address gaps on a per-metabolite basis and can provide multiple solutions but cannot avoid thermodynamically infeasible cycles (TICs), invariably requiring lengthy manual curation. To address these limitations, this work introduces an optimization-based multi-step method named OptFill, which performs TIC-avoiding whole-model gapfilling. We applied OptFill to three fictional prokaryotic models of increasing sizes and to a published GSM of Escherichia coli, iJR904, and to the creation of a novel model of the polyextremotolerant fungus *Exophiala dermatitidis*. These applications resulted in holistic and infeasible cycle-free gapfilling solutions. In addition, OptFill can be adapted to automate inherent TICs identification in any GSM. Overall, OptFill can address critical issues in automated development of high-quality GSMs. The OptFill tool is a multi-part optimization-based algorithm designed for conservative reconstruction of stoichiometric models of metabolism. The original OptFill tool (the first mentioned publication in the chapter header), had certain limitations of speed, size, and accuracy, which was addressed in a later application and publication of this tool. This chapter then, will be a synthesis of the former publication with the algorithmic enhancements from the second publication.

2.2. INTRODUCTION

The use of systems biology in uni- and multi-cellular organisms (e.g. plants and animals) to engineer or enhance desirable phenotypes and study system-wide metabolic processes is well-established and capable of affecting the lives of millions of individuals, such as in the case of artemisinin production in yeast or enhancing the nutritional value of agricultural products (Beyer et al., 2002) (Hall et al., 2008). As opposed to traditional qualitative approaches, computational approaches based on stoichiometric Genome-Scale Models (GSMs) of metabolism can be used to predict non-intuitive genetic interventions (Srinivasan, Cluett, & Mahadevan, 2015) by accounting for gene-protein-reaction (GPR) links. GSMs may also lead to increased understanding of how a change in environment, a change in organism nutrition, or a gene knockout, can affect the entire metabolic system of an organism through tools such as Flux Balance Analysis (FBA) (Orth et al., 2010), OptKnock (Burgard, Pharkya, & Maranas, 2003), and OptForce (Ranganathan, Suthers, & Maranas, 2010). GSMs have been developed for many prokaryotic (Magnúsdóttir et al., 2016)(Shoaie et al., 2013), animal (Brunk et al., 2018), plant (Gomes de Oliveira Dal'Molin et al., 2015)(Saha, Suthers, & Maranas, 2011), and fungal (Andersen, Nielsen, & Nielsen, 2008)(J. Liu, Gao, Xu, & Liu, 2013) systems, enhancing mechanistic understanding and exploration of system-wide metabolism in such organisms as *E. coli* (Ranganathan et al., 2010), cyanobacteria (Saha et al., 2016), yeast (Ng, Jung, Lee, & Oh, 2012), and other species (Saha et al., 2011)(S. Gudmundsson et al., 2017)(Shoaie et al., 2013)(Islam, Al-Siyabi, Saha, & Obata, 2018). GSMs are typically reconstructed by gleaning information on gene annotations, enzyme functions, associated reactions, and reaction directionality from major public databases such as KEGG (Kanehisa, Furumichi, Tanabe, Sato, & Morishima, 2017), ModelSEED (Overbeek et al., 2005), the NCBI (Limviphuvadh et al., 2018), MetaCyc (R. Caspi, 2006), K-Base (Arkin et al., 2018), and BIGG (King et al., 2016). At present, there is no complete knowledge of any genome. For instance, the annotated genome of one of the most prolifically studied organisms, *Escherichia coli* strain K-12 substrain MG1655, contains about 6.8% putative proteins and 16.1% uncharacterized proteins (UniProtKB, 2018). Furthermore, approximately 61% of proteins lack an Enzyme Commission (EC) number, which is important for the identification of GPR links in any GSM reconstruction (UniProtKB, 2018). Inevitably, incomplete gene annotation and system knowledge (including reaction direction) leaves metabolic gaps, imbalances, or Thermodynamically Infeasible Cycles (TICs) in any initial GSM reconstructions, leaving the model incomplete. Particularly problematic are TICs, sets of reactions which can carry flux in the absence of nutrition provided to the model because their net stoichiometry is zero, also known as futile cycles or type III reactions (Thiele & Palsson, 2010). These cycles can negate metabolic costs (Thiele & Palsson, 2010), report infeasibly large reaction rates, be difficult to identify (De Martino, Capuani, Mori, De Martino, & Marinari, 2013)(Schellenberger, Lewis, & Palsson, 2011), and inhibit the proper function of optimization-based tools which rely on duality to optimize multiple objectives such as OptKnock (Burgard et al., 2003) and OptForce (Ranganathan et al., 2010).

A significant challenge to reconstruct GSMs is the amount of time and manual labor required to curate these incomplete reconstructed models, addressing various issues such as element and charge balances; reaction directionality; metabolic gaps; TICs; and other inconsistencies. Hence, it often requires months to years of manpower before a predictive model is generated (Thiele & Palsson, 2010), which is a prerequisite for conducting research on phenotypic enhancement or study metabolism. Two of the most challenging aspects of model development are the identification and elimination of TICs, as well as the resolving of metabolic gaps.

The existing methods/tools that have been developed to address the identification and resolution of TICs can be broadly categorized into four groups: i) methods that can identify existing TICs in a model (De Martino et al., 2013), ii) methods that can force no-flux through existing TICs in a model (Schellenberger et al., 2011)(Nigam & Liang, 2007)(Chan, Wang, Dash, & Maranas, 2018), iii) a combination of the previous two (Chan et al., 2018), and iv) methods eliminating TICs by manipulating the metabolic network. Although developing these is a significant step toward building a better and more predictive GSM, there remain challenges that need to be addressed. For the first approach, Monte Carlo sampling-based method (De Martino et al., 2013) cannot guarantee the identification of all TICs as it is a stochastic approach. The second approach is the avoidance of TICs by the application of Kirchoff's Loop Law in methods such as Loopless COBRA (Schellenberger et al., 2011). This approach does successfully avoid TICs, but does not address the root cause in the model which can make some models problematic for tools such as OptForce which require no TICs (Ranganathan et al., 2010). Another approach is the addition of thermodynamic constraints to the model using known thermodynamic quantities (Nigam & Liang, 2007), which works well for well-studied organisms for which these *in vivo* parameters are known, but is more difficult to implement for non-model organisms. The third approach that combines these two approaches, such as the one demonstrated by Chan et al. (2018), has shown promise and computational tractability. However, this has generally been employed as a set of loopless constraints, rather than as a method to avoid the inclusion of TICs in gapfilling. The fourth method has been used to address TICs in energy metabolism, which can allow the model to produce unlimited energy severely hampering model accuracy, by applying a variation of optimization-based tool GLOBALFIT (Fritzemeier, Hartleb, Szappanos, Papp, & Lercher, 2017). GLOBALFIT has been used by Fritzemeier et al. (2017) to identify the minimal network changes to address erroneous energy cycling in metabolic network models. These changes could take the form of removal of reactions and/or correcting of reaction direction and address root causes of TICs without using loopless constraints when applying *in silico* analysis tools.

It should be noted that not all the cycles in biological systems are infeasible cycles. Some cycles, such as the Calvin cycle or the citric acid cycle are well-known biological cycles. These differ from infeasible cycles in that one cycle has some net effect, in the case of the Calvin cycle this net effect of each revolution is to fix carbon dioxide to a sugar by expending cellular energy. In contrast, thermodynamically infeasible cycles result in no net production or consumption per each revolution. It should also be noted that some reactions do proceed in both directions at the same time in the same subcellular compartment in a cell, with their relative rates limited by thermodynamic considerations. While some models do include *in vivo* thermodynamic information, the precise value, or more often range of values, for the Gibbs free energy and other important thermodynamic properties of a reaction are often unknown aside from being able to specify reaction direction (Thiele & Palsson, 2010). Therefore, for all but the best-studied organisms, imposing thermodynamics-based limitations on reaction rates to preclude thermodynamic cycling is very difficult if not impossible.

To address and resolve metabolic network reconstruction gaps, GapFind and GapFill (Satish Kumar, Dasika, & Maranas, 2007) are some of the most common tools used (Pitkänen et al., 2014)(Henry et al., 2010)(T. Y. Kim et al., 2012). GapFind and GapFill are optimization-based Mixed Integer Linear Programming (MILP) problems, and have been successfully implemented in the reconstruction of metabolic models, of prokaryotic and eukaryotic biological systems such as cyanobacteria *(Synechocystis* sp. PCC 6803 and *Cyanothece sp ATCC 51142)* (Saha et al., 2012), corn (*Zea mays*) (Simons et al., 2014), yeast (*Saccharomyces cerevisiae*), and Chinese hamster ovary cells (Chowdhury, Chowdhury, & Maranas, 2015). Other methods of automated gapfilling which build on the capabilities of GapFill include GenDev (Latendresse & Karp, 2018), FastDev (Latendresse & Karp, 2018), likelihood-based gapfilling (Karp, Weaver, & Latendresse, 2018), and phenotype-based gapfilling (Cuevas et al., 2019). All these tools are constructed with the aim of increasing the accuracy of the GapFilling method, through comparison to some level of data such as phylogenetic, phenotypic, or genetic. In this work, a problematic aspect of all these tools is considered which these other tools were not built to address. Despite their success, the tools for gapfilling have significant limitations including: i) gaps are addressed on a per-metabolite basis (as opposed to a whole-model holistic approach), ii) thermodynamic feasibility is often not considered, and iii) reaction direction is not considered in gapfilling, rather all reactions are added reversibly. From the first and second limitations, several problems arise including: i) inability to guarantee that the minimum number of reactions are added to fix metabolic gaps on a whole-model basis; ii) inability to identify and avoid unfavorable interactions between multiple gap fixes (often, TICs); and iii) differences in the resultant model dependent of the individual curator.

To address current TIC-finding and gapfilling method limitations, this work introduces a multi-step optimization-based MILP method. The first step is to solve an iterative optimization-based TIC-Finding Problem (TFP) which identifies potential TICs, which may be caused by adding reactions from a database in a given direction (see Figure 2.1). This method uses optimization and binary variables as opposed to null space matrices used by other methods which identify reactions participating in TICs (Saa & Nielsen, 2016) or TICs (Chan et al., 2018), and thus can provide a greater level of detail for each inherent or potential TIC. This problem is unique as it considers the direction and relative flux rate of reactions participating in TICs and can be easily adapted for the purposes of model curation sans database for the resolution of inherent TICs. The second step involves the solving of three optimization-based problem, the Connecting Problems (CPs), which are highly similar but have different objectives. The first Connecting Problem (CP1) is the maximization of model metabolites successfully connected to metabolic network, e.g. maximizing the number of metabolites which the connected model can now produce, while avoiding the addition of TICs. The second Connecting Problem (CP2) is the minimization of the number of reactions required to achieve the objective of CP1. The third connecting problem (CP3) is the maximization of the number of reactions to be added reversibly from the database to achieve the objectives of CP1 and CP2 subject to avoiding TICs. The connecting problems are unique in that, unlike other gapfilling algorithms, CP solutions provide whole model gapfilling solutions guaranteeing the minimum number of reactions being added for the maximum number of fixed metabolites. As proof of concept, the OptFill approach is applied to three test stoichiometric models of increasing sizes (models of 28 to 210 reactions, databases of 17 to 77 reactions) with designed metabolic gaps, and one smaller (1074 reactions) GSM of *Escherichia coli* with acknowledged metabolic gaps (Reed, Vo, Schilling, & Palsson, 2003) using another GSM of *E. coli* as the basis for a database (Feist et al., 2007). With the computational resources at hand, the full OptFill method is limited to relatively smaller stoichiometric models and databases but should be applicable to larger models and databases given access to greater computational power.

2.3. RESULTS

2.3.1. Development of OptFill

OptFill was conceived and developed to address the limitations of the current state-of-the-art GapFind/GapFill (Satish Kumar et al., 2007) tool. The initial stages of the design-build-test (DBT) cycle contained the first Test Model (TM1) and the first Test Database (TDb1) and involved only a single connecting problem. TM1 was constructed as a small stoichiometric model involving starch and glycolysis metabolism to produce ethanol but with metabolic gaps preventing growth (see Figure 2.1). TDb1 was designed to have the capacity to fill these gaps, at the expense of potentially producing TICs. In the DBT cycle, it was soon realized that the TFP was necessary to define the potential TICs which might occur. The TFP was built to solve for the smallest TICs (i.e., the TICs with the smallest number of participant reactions) first and then solve for larger TICs to prevent multiple TICs masquerading as a single TIC solution. The workflow representing the TFP is shown in Figure 2.2. The CPs were developed to ensure consistency in the number, order, and identity of the CP solutions while avoiding the addition of the whole set of TICs identified as potentially occurring between the model and database. See Figure 2.3 for the conceptual formulation of each type of problem. All problems which are part of the OptFill tool are Mixed Integer Linear Programming (MILP) problems which ensure global optimality of each solution in each iteration.

On occasion, the feasibility constraints used might be too strict to return a feasible solution to the CP problems which could result in execution errors prematurely ending OptFill before completion. Therefore, an error handling framework was built around each CP problem allowing a one-time relaxation of feasibility constraints. These frameworks are shown in Figure 2.2. OptFill is ended when CP1 no longer has a feasible solution even when feasibility constraints are relaxed (which occurs because previous solutions are prevented from being re-identified) since at that point none of the CP2 and CP3 will have a feasible solution. Further, all OptFill runs described used non-standard CPLEX solver options, which effectively eliminated most types of cuts. This caused some level of reduction to the solution space, particularly those which could result in non-optimal solutions being reported as optimal. These included flow, zero-half, and Gomory fractional cuts among others. This was done because the order of solutions is important in the OptFill method, and the order of solutions also has bearing on the number of solutions returned. See Transparent Methods for further detail.

2.3.2. Application of OptFill to Test Models

After finalizing the formulation (see Figure 2.3 and Transparent Methods) and workflow (Figure 2.2) of OptFill, a detailed analysis of OptFill results with respect to TM1 and TDb1 was undertaken. Some qualitative results of the application of the OptFill workflow to TM1/TDb1 are shown in Figure 2.1, which include the initial model and database, Figure 2.1(A); the combination of the model and database, Figure 1(B); selected identified potential TICs, Figure 2.1(B); and selected identified CPs' solutions, Figure 2.1(C). As is shown in Figure 2.1(A), TM1 is too disconnected to produce biomass, but in combination with TDb1 can potentially produce biomass. When the TFP is applied (Figure 2.1(B)), 31 potential TICs consisting of 3 to 12 reactions (hereafter, sizes 3 to 12) were identified. The average solution time (when a solution was found) was 0.175 s (σ=0.0727 s, min=0.0870 s, max=0.378 s). It should be noted that all solve times reported here are not constant, even if using same resources. Figure 2.1(B) highlights 5 potential TICs which were identified. The first two TICs identified, TIC #1 and #2, show that the TFP can identify TICs occurring only in the database; that TICs consisting of the same metabolites and reactions are identified separately if reaction directions are different; and that two of the smallest TICs are identified. Potential TIC #9 shows a TIC of moderate size (for TM1/TDb1) which contains an irreversible model reaction related to Non-Growth Associated Maintenance (NGAM) and, therefore, will not have a companion potential TIC of opposite direction, unlike potential TIC #1 and TIC #2 (in the opposite direction). Further, this highlights the potential for infeasible cycling which effectively negates the cost of NGAM of the model. If added in its entirety, NGAM would be irrelevant to the model at any value and would significantly reduce model accuracy. This TIC might not be manually identified since NGAM is usually a fixed quantity. Potential TIC #10 highlights another type of infeasible cycling involving ADP/ATP, but this cycling essentially negates the cost of phosphorylation/ dephosphorylation of glucose-6-phosphate isomers. Finally, potential TIC #31 is included to highlight a non-intuitive TIC, in addition to be the largest TIC identified. This TIC involves the separate cycling of sugars and 3-carbon molecules linked and is made possible by ADP/ATP cycling (sugar cycling consumes ATP and 3-carbon cycling produces ATP). These examples illustrate that many, but not all, potential TICs involve the infeasible cycling of energy molecules, which should be particularly avoided in the reconstruction of models of metabolism as this can result in negated costs for various biological activities with which a cost should be associated. This negated cost can often result in increased model growth rate and reaction fluxes, reducing the model's accuracy.

The model, database, and TFP solutions form the input for the CPs. Before solving the CPs, a modified version of CP1 was run which prohibited the addition of database reactions. This modified CP1 reported that the raw TM3 model was capable of producing no metabolites. The CPs, when applied to TM1 and TDb1, identified 24 potential solutions which connected between 31 and 33 metabolites with the additions of 6 to 10 reactions, of which 0 to 6 could be reversible without TICs. The average time to solve all three CPs for each solution was 0.639 s (σ=0.147 s, min=0.433 s, max=0.950 s), see Figure 2.4. From the FBA performed on each connecting problem solution with the objective of maximization of biomass, the mean maximum biomass production rate of the set of connected models was 2.43 h^{-1} (σ=0.394 h^{-1}, min=1.44 h^{-1}, max=2.90 h^{-1}). Solution times for the FBA code were not recorded as FBA solution time is generally low. Two connecting problem solutions, the first and the last, are shown in Figure 2.1(C). These solutions are notably different in terms of the number of model metabolites connected by the CPs' solution (green boxes in the metabolic sketch), the number of intermediate metabolites introduced by these solutions (yellow boxes), the number of database reactions introduced (orange arrows), and even the use of energy molecules. For instance, CPs' solution 1 introduces only two additional metabolites and 6 reactions reversibly from the database which are part of the CPs' solution and connects all but two model metabolites. The first is acetate, which is a dead-end metabolite. The second is the extracellular proton, which suggests that the model is small enough that all protons produced are also consumed. This solution has the slowest growth rate of all connecting problem solutions. On the other hand, CPs' solution 24 connects two fewer metabolites than CPs' solution 1, requires two more reactions, introduces two more intermediate metabolites, and has a higher growth rate. It is hypothesized that this is due to the more efficient production of ATP allowed by reaction R01512[c] (enzyme ATP:3-phospho-D-glycerate 1-phosphotransferace in the cytosol), which is present in many other high-biomass solutions. This reaction allows two dephosphorylation events to produce ATP, as opposed to only one (the other event occurring by hydrolysis).

Two larger test models were built next to study the increase in number of solutions and time required to reach those solutions by OptFill and, ultimately, to investigate its scale-up potential. Each test model was built from an OptFill solution of a previous solution to highlight the ability of this tool to be applied in sequence. In the application of OptFill to the existing models of organisms, careful attention must be paid in selection of a CPs' solution to accept, including considerations of energy metabolism, predicted growth rates, and remaining unconnected metabolites. Here, CPs' solution 1 was selected and combined with TM1 as the base of the second test model (TM2). Reactions and metabolites from the fatty acid biosynthesis and the pentose phosphate pathway were added to this base, in which gaps were manually created. Reactions which could address these gaps formed the second test database (TDb2). Redundant metabolic functions were added to TDb2 to allow for potential TICs. Similarly, the third test model (TM3) was built from the first CPs' solution of TM2 and TDb2. Additionally, a bank of reactions from the amino acid synthesis pathways, including redundant functionalities, was created. This bank was automatically (randomly) sorted between those reactions which would be added to complete TM3 (~80% of bank reactions) and those which would constitute the third test database (TDb3, ~20% of bank reactions). As random sorting was used, a modified version of the TIC-Finding Problem (modified TIC-Finding Problem, mTFP), was used to identify inherent TICs in TM3 and TDb3 which resulted from the random assortment of the bank reactions. The reactions most-commonly participating in identified inherent TM3 TICs were moved to the TDb3 until no inherent TICs remained (5 reactions in total).

For OptFilling of TM2/TDb2, 51 TICs consisting of 3 to 26 reactions were identified by the TFP, with a mean solution time of 0.131 s (σ=0.0405 s, min=0.0850 s, max=0.308 s). The largest TIC, potential TIC #51 consisting of 26 reactions, would have largely been very difficult to identify by a non-automated method as it spans six KEGG pathways including glycolysis; the pentose phosphate pathway; purine metabolism; nicotinate and nicotinamide metabolism; starch and sucrose metabolism; riboflavin metabolism. TIC #51 involves the cycling of 3-, 4-, 5-, and 6-carbon molecules, energy molecules (ATP, NADH, and NADPH), and energy molecule hydrolysis. This TIC can be found in GitHub OptFill or Mendeley Data repositories accompanying this work.

Before solving the CPs, the modified CP1 was run and reported that the raw TM2 model was capable of producing no metabolites. Fifteen potential CPs' solutions were identified which each connected 90 to 94 metabolites with the addition of 17 to 23 reactions, of which 0 to 19 could be reversible without TICs. The average time to solve all three CPs for each solution was 1.40 s (σ=0.639 s, min=0.404 s, max=2.65 s), see Figure 2.4. From the FBA performed, the biomass production rate of most CPs' solutions applied to TM2 was 1.31 h^{-1}, for 10 solutions, and 1.36 h^{-1} for the remaining five. In the CPs' solutions, those with the highest biomass have fewer metabolites which could be connected (all solutions with higher biomass production were generated after lower biomass production solutions). Those with the higher biomass production rates generally have one fewer reaction which requires ATP hydrolysis, and therefore has slightly more energy in the system to spend on the production of biomass than their lower biomass counterparts.

Similarly, OptFill applied to TM3/TDb3 resulted in the identification of 60 TICs consisting of 3 to 31 reactions by the TFP and 177 potential CPs' solutions which each connected 202 to 214 metabolites with 12 to 17 reactions, of which 1 to 12 could be reversible without TICs. As earlier, the modified CP1 was used to identify 54 metabolites which the raw TM3 was capable of producing. The mean TFP solution time was 0.240 s (σ=0.0756 s, min=0.141 s, max=0.541 s), whereas the mean CPs' solution time was 0.985 s (σ=0.249 s, min= 0.573 s, max=1.86 s), From the FBA performed, the mean biomass production rate of the connected model was 3.29 h^{-1} (σ=0.179 h^{-1}, min=3.11 h^{-1}, max=3.47 h^{-1}). Runtime and solution metrics for all solutions are shown in Figure 2.4. Unlike TM1 and TM2 OptFilling results, there was no solution where all database reactions to be added by the CPs' solution could be added reversibly. This indicates that, for all solutions, the direction in which database reactions are added is important to avoid TICs to produce a model without the disadvantages of TICs described previously. Furthermore, the biomass production rate does not appear as dependent on either the number of metabolites connected or reactions added as in previous CPs' solution sets. Instead, the biomass production rate seems to most depend on the method of sulfate assimilation.

2.3.3. Application of OptFill to *i*JR904

In order to show how the OptFill workflow might scale up to a GSM, the *i*JR904 model of *Escherichia coli* consisting of 761 metabolites, 1074 reactions, and 904 genes (Reed et al., 2003) was selected as the base model to fix. The *i*AF1260 model, a model extending onto *i*JR904, consisting of 1598 metabolites, 2,381 reactions, and 1260 genes (Feist et al., 2007) was selected to serve as the set of reactions from which to build the database. *i*JR904 contains 70 dead-end metabolites (Reed et al., 2003) which need fixing. Before applying OptFill, some minor formatting changes were made (described in Transparent Method and in the related GitHub OptFill or Mendeley Data repositories accompanying this work) and it was decided that carbon-limited aerobic growth using acetate would be the condition for which *i*JR904 model would be fixed. Metabolite exchange rates were taken from Reed *et al.*, 2003 to describe this growth condition.

In order to create the database which would be applied to *i*JR904, all *i*AF1260 exchange reactions and reactions with names identical to those in *i*JR904 (which were assumed to be the same reaction as the former was built from the latter) were removed from *i*AF1260 to form the initial database which consisted of 1441 reactions. This proved too computationally intensive for the resources, and therefore this database was further simplified in a manner which it is suggested others with limited computational resources might also use. First, the *i*AF1260-based database and *i*JR904 were combined in single model file and Flux Variability Analysis (FVA) (Steinn Gudmundsson & Thiele, 2010) was performed (see Table S1, see section 7.2 for how to access this file, iJR904, Related to Figure 2.4). Those *i*AF1260 reactions capable of holding flux as determined by FVA (715 reactions) were defined as the database of functionalities to be used with OptFill.

OptFill was performed on *i*JR904 using this database. This still resulted in a slow OptFill process, therefore solutions which were reported (i.e., 4 identified) in the allotted solve time of 24 hours were collected. All *i*AF1260 reactions which participated in at least one solution (a total of 182 reactions) were selected as the basis of the third *i*AF1260-based database. This resulted in significantly lower computational requirements for the application of OptFill. This database was found, upon application of OptFill, to be without TICs. For the purposes of demonstration and showing how the increase of TFP solution time changes with model and database size, it was arbitrarily decided to add six reactions manually from the previous database which could participate in potential TICs between the model and database, but which did not create TICs only within the database. Further, the mTFP was applied to the *i*JR904 model. From the mTFP results, it was noticed that in *i*JR904, some reactions were included in the model twice, both as reversible and irreversible, causing inherent TICs in the *i*JR904 model involving these duplicate reactions. It was decided to move the irreversible reactions of each duplicate pair to the database (nine reactions in total) so that all *i*JR904 models were still present in the OptFill in some capacity. The final *i*AF1260-based database for the OptFilling of *i*JR904 totals 188 reactions. Initial, final, and intermediate iAF1260-based databases used can be found in Table S1. iJR904, Related to Figure 2.4. iJR904 and in the GitHub OptFill or Mendeley Data repositories accompanying this work.

Demonstrated here is a procedure by which the database applied to a model can be significantly decreased in size to reduce computational cost of the OptFill method, while still effectively addressing metabolic gaps. This can be summarized as i) eliminate all duplicate reactions; ii) perform FVA on a pseudomodel which is a combination of the database and model and use the results to eliminate reactions which cannot carry flux; and iii) perform OptFill using databases with larger solution time, collect a few sample solutions, and use the set of reactions participating in sampled solutions as the database. Applications of steps i) and ii) as well as iterative applications of iii) might be used by modelers to shrink the databased used in OptFilling to a size which is possible to solve in a modest period of time given the computational resources available.

This final *i*AF1260-based database was used to OptFill *i*JR904 model. In this final iteration, there were 25 TICs of size 2 to 8 reactions identified. The associated mean TFP solution time was 0.410 s (σ=0.0978 s, min= 0.330 s, max=0.687 s). The TICs identified were generally simple, as they stem from reactions manually added to the database which cause TICs. Eleven TICs occur between just two reactions, and a further 4 involving only a single database reaction. Each of these effectively precluded a single database reaction from being added in a certain direction. When the CPs were applied to *i*JR904, it was found that the CPs' solution time had increased considerably from that of other models, to a mean of 236 s (σ=329 s, min= 15.3 s, max=1010 s). The solution time of this model was significantly increased due to disabling of many types of cuts which a solver might use to decrease solution time, but which lead to non-optimal solutions being reported as optimal. These are particularly relevant because minor cuts, such as those that accept a 0.5% reduction in the optimal solution value, can change the number of metabolites connected by the CPs by two or more for GSMs. As the order of solutions is important, even these minor relaxations were deemed problematic and were therefore mostly disabled, leading to increased solution time. If these cuts were allowed, CPs' solution time would have been approximately an order of magnitude less than reported here. The modified CP1 problem reported that the *i*JR904 model was capable of producing 358 metabolites under the given aerobic growth on acetate conditions, and all CPs' solutions connected 418 metabolites with the addition of 86 reactions. All CPs' solutions produced biomass at a rate of 0.108 h^{-1}. This is likely a result of the database reduction steps taken. The variation on the CPs' solution occurred in the number of connecting reactions which could be added reversibly, ranging from 5 to 86. The full set of solutions can be found in the GitHub OptFill or Mendeley Data repositories accompanying this work. It can be seen in Figure 2.4(I) that efforts to prevent non-optimal solutions from being reported as optimal were not entirely successful. There exists one CPs' solution, solution #72, where the optimal (maximum) CP3 solution value is 5, whereas the optimal (maximum) CP3 solution value was 11 from solutions #71 and #73. This occurred when all solutions were subject to approximately the same constraints (save the integer cuts necessary to prevent repeated solutions). It is noted earlier that many types of cuts were disabled, but not all, and one type of cut or other solver setting allowed this non-optimal solution to be reported as optimal, however; eliminating all such cuts and settings proved prohibitively time-consuming. Therefore, the settings, which can be found in the GitHub OptFill or Mendeley Data repositories accompanying this work, were selected as those which, for this work, best balanced solution order and solution time.

In the OptFilling solutions of *i*JR904, several trends can be noticed which were not present in the smaller test models. First, when performing FBA, with the objective of maximizing biomass, on the resultant OptFilled *i*JR904 model, not all reactions from the database held flux when biomass was maximized. This is because these reactions make it possible for the model to produce metabolites which are not required for the production of biomass or provides an alternative pathway for the production of biomass which might be less efficient. This does not mean that these connected metabolites are unimportant under other, equally valid, objective functions, for instance the connected metabolites may be bioproduction targets. Further, some TICs exist between *i*JR904 model reactions in the OptFill solutions and notably one database reaction. For most model reactions, these TICs occur because forward and reverse reactions are written separately. The TIC involving the database reaction resulted from the proton uptake exchange reaction being allowed a very high reaction flux in the *i*JR904 model. The TFP was performed with all exchange reactions fixed to a flux of zero, therefore the TFP did not identify this TIC which involved an exchange reaction. When the exchange reactions were allowed to carry flux again in the CPs, the high proton uptake rate (here, 1000 mmol/gDW·h) allowed the cycling of reactions. These resulting TICs highlight two important considerations in using OptFill. First, the mTFP should be used in combination with manual editing of the model to ensure that the model does not contain inherent TICs as the usual OptFill workflow will not address inherent TICs. Second, reasonable bounds should be applied to all exchange reactions (such as the proton uptake reaction) and to forward and reverse reaction pairs to prevent TICs in the OptFilled model.

2.3.4. OptFill Solution Times

With the caveats of the available resources (see Transparent Methods for information on the software and hardware tools available for this work), the TFP seems to have a per-TIC average solution time with linear dependence ($R^2 \geq 0.89$) on size of model and/or database used, see Figure 2.4(A) through (C). The same procedure was applied to the aggregated CPs' solution time, but with significantly different results. Exponential trend lines were able to fit with a high correlation coefficient ($R^2 \geq 0.96$) between model, database, total system size, and CPs aggregated solution time. This is indicative of a strong correlation between CPs aggregate solution time number of reactions in the total system, and that increasing total system reactions greatly increases CPs aggregate solution time.

2.4. DISCUSSION

Introduced here is an optimization-based tool, OptFill, which can be used to increase the automation of the curation of GSMs. This tool can either be used to automate the filling of metabolic gaps in a reconstructed model or to automate the identification of TICs for manual resolution (via mTFP). In this work, the OptFill was applied in sequence to three test models of increasing size as well as to a GSM of *E. coli*, *i*JR904. These applications combined with some solutions for holistically gapfilling metabolic models, the computational expense of the tool, and a method for reducing that expense highlighted the utility of OptFill.

This method has considerable potential to be adapted to other metabolic systems (both eukaryotic and prokaryotic) and is not specific to any identifier system such as KEGG or ModelSeed. For instance, while all test models as well as *i*JR904/*i*AF1260 have been prokaryotic systems, there is no reason why this approach would not similarly work in a eukaryotic organism. Further, the framework is flexible enough that any system of reaction and metabolite identifiers, such as KEGG (Kanehisa et al., 2017), MetaCyc (Ron Caspi et al., 2014), BIGG (King et al., 2016), K-Base (Arkin et al., 2018), or custom identifiers, may be used for metabolites and/or reactions, making this tool applicable to a wide variety of existing GSM-building methods. This was demonstrated in this work as KEGG identifiers were used in the test models, whereas BIGG identifiers were used by the *i*JR904 and *i*AF1260 models (Reed et al., 2003; Feist et al., 2007).

From the observation of TFP solution times, it is evident that the TFP and mTFP could scale-up to genome-scale models of metabolism as a linear trend line ($R^2 \geq 0.89$) strongly describes the per-TIC solution time given the computational resources at hand. So long as the number of TICs in the system remains reasonable, this portion of OptFill is transferrable to large-scale GSM systems, or to situations where computational resources are limited. The transferability of the OptFill method is likely limited by the computational resources available to the end-user, as the aggregate solution time of the three CPs is well-described by an exponential trend line ($R^2 \geq 0.97$). This suggests that those without access to powerful computational resources may have difficulty implementing OptFill in a reasonable timeframe, unless, for instance the end-user makes trade-offs between the solution order (e.g. each subsequent solution is truly globally optimal) and solution time. These trade-off issues, such as shown in a minor way with the OptFilling of *i*JR904, may likely be fixed by more advanced MILP solvers which are currently available or by advances in optimization which may be made in future.

When implementing OptFill in other systems, a high-quality model and database should be used in order to limit both the number of solutions and the time the OptFill method takes to complete. This is primarily due to the number of feasible and unique combinations possible. For instance, if a multi-step reaction is included in a database in addition to its component reaction steps, this can potentially double the number of solutions found by both the TFP and CPs. To explain, if the multi-step reaction participates in *n* TICs, then its component step reactions would participate in *n* TICs. This results in *2n* TICs, where only *n* TICs need be identified. The same argument applies for CPs' solutions. This error in model reconstruction could then double (or more) the number of TICs and CPs' solutions as well as the total OptFill runtime in a stroke. In larger models, such issues can result in a significant expenditure of time (potentially days) and computational resources which need not be expended should the model and database used to be of high quality. Such an issue is elsewhere referred as a combinatorial explosion (Burgard et al., 2003).

This was shown in this work in the failure to achieve a reasonable number of solutions or reasonable solution times in the OptFilling of *i*JR904 with a poorly curated database based on *i*AF1260; however, when the database was better curated, reasonable numbers of solutions and solution times were achieved. Therefore, it is important to address as many inherent TICs which occur both in the model and in the database as feasible using the mTFP on both the model and the database to identify and address these TICs.

While throughout this text reaction cycling in the absence of nutrition (i.e., Thermodynamically Infeasible Cycling) is described as a phenomenon which is to be avoided in GSMs, this is not always the case. In many biological systems, cycling of some type does occur and the absence of that cycling in the models might affect their accuracy. However, cycles included in a GSM should be carefully considered with respect to their biological relevance, magnitude, and effect, particularly when they occur in the absence of nutrition provided to the model. In essence, this work can be used to remove and/or avoid all cycling which can occur in the absence of nutrition provided to the model or to ensure that cycles retained are deliberate and have biological relevance if included. If cycles occur in a GSM model in the absence of nutrition provided to the model and are biologically relevant, best practice should be to use other literature data available to limit the scope of the cycling to feasible number. This trade-off must be considered when applying the OptFill algorithm, or when choosing to use some type of algorithm which employs the loopless constraints.

This is the essential difference between what is proposed here as the OptFill tool and other algorithms such as the algorithm employed by Chan et al. (2018) to identify all TICs in a model and avoid them. The TIC finding portions of the algorithm are largely equivalent, although Chan et al. (2018) may identify TICs faster. OptFill then precludes these TICs from being added as part of a gapfilling solution so that the resultant reconstructed metabolic model contains no inherent

TICs. However, Chan et al. (2018) accepts these TICs in the reconstructed network and seeks to limit flux through these TICs so that the resulting model fluxes are feasible. The OptFill approach presents an alternative to the need to use loopless algorithms on the gapfilled model and allows use of algorithms which are sensitive to the presence of TICs, such as OptForce (Burgard et al., 2003)(Chan et al., 2018) without modifying these algorithms for the use of various loopless algorithms which may be computationally expensive.

In future, this work will be used as a gapfilling and curation strategy for the development of GSMs of any prokaryotic and eukaryotic systems. In concert with advances in optimization solvers and available computational resources, these methods (i.e., the TFP, CPs, and their modified versions) will provide an alternative holistic method of model curation. At present, those model-building tools with high computational power at their disposal, such as ModelSeed (Overbeek et al., 2005) and K-Base (Arkin et al., 2018), may well be able to implement OptFill and its components for large GSMs to improve their automated curation capabilities. In addition, with the available computational resources and some adjustments (as explained earlier), Optfill is being implemented to improve the connectivity and predictive capability of the GSM of a non-model purple non-sulphur bacterium (Alsiyabi, Immethun, & Saha, 2019) and to develop the GSM of a melanized fungal strain.

2.5. FIGURES

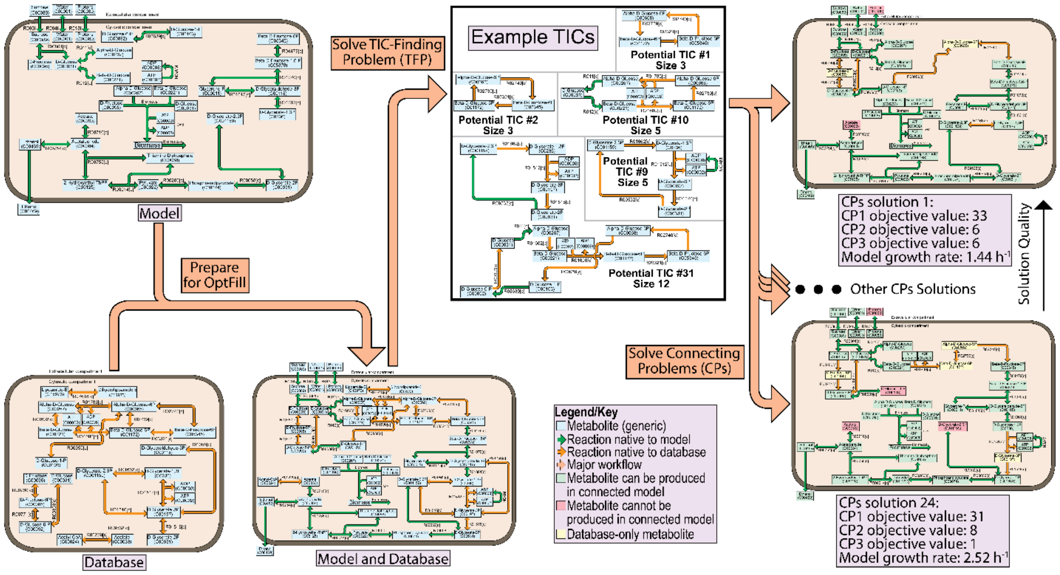


Figure 2.1: Conceptual drawing of how the OptFill algorithm works.

Extended Caption: Background colors of this visualization correspond to the workflow presented in Figure 2.2, where the colors green, blue, and purple correspond to preparation for OptFill; the TFP and its framework; and the CPs and its framework, respectively in both images. A) Shows that the model and database are separate, but are both used in the workflow to prepare for OptFill and in OptFill itself. B) Shows how the combined model and database might appear, and how this combination is used in the TIC-Finding Problem to identify potential TICs which might occur between the model and the database. Selected identified potential TICs are shown here as illustrative examples. Potential TICs #1 and #2 illustrate how TICs occurring in different directions are identified as separate TICs, how identified TICs might only occur between database reactions, and illustrate the two of the smallest identified TICs. Potential TICs #9 illustrates a larger TIC which makes use of an irreversible reaction (NGAM), and therefore has no opposite-direction TIC, making the direction of the other reactions important. Potential TICs #10 and #31 illustrate infeasible cycling involving an energy molecule (ADP/ATP), in addition to potential TIC #31 being the largest identified TIC. C) Show the application of the Connecting Problems (CPs) and the first and last solution of the CPs. These solutions differ in the number of model metabolites which could not be connected (red boxes); the number of metabolites introduced to the model (yellow boxes); the number and reversibility of database reactions added (orange arrows); and the resultant model growth rate.

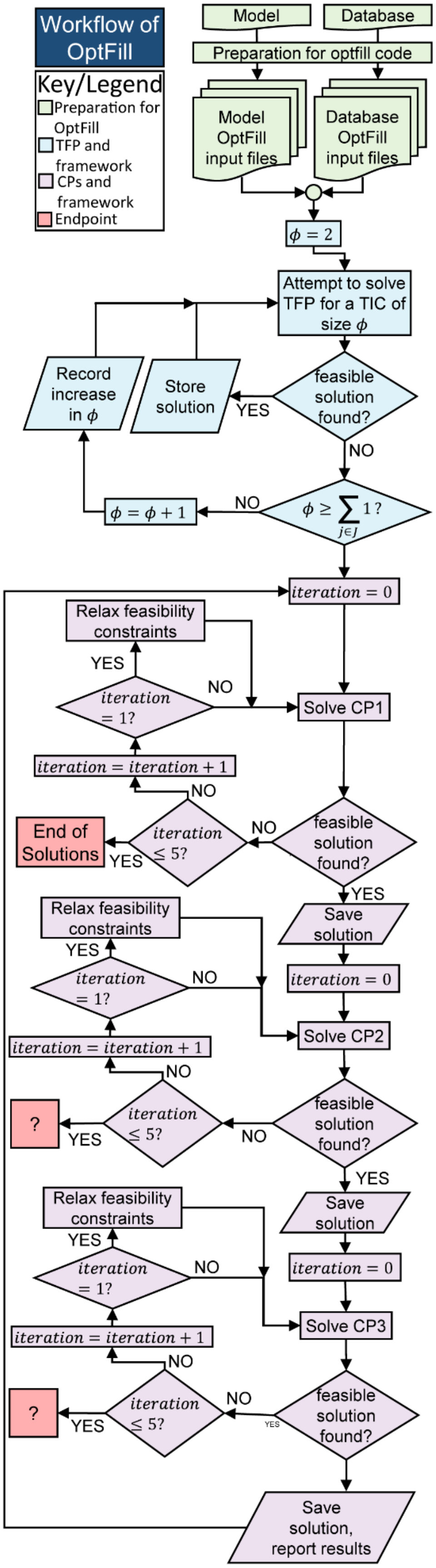
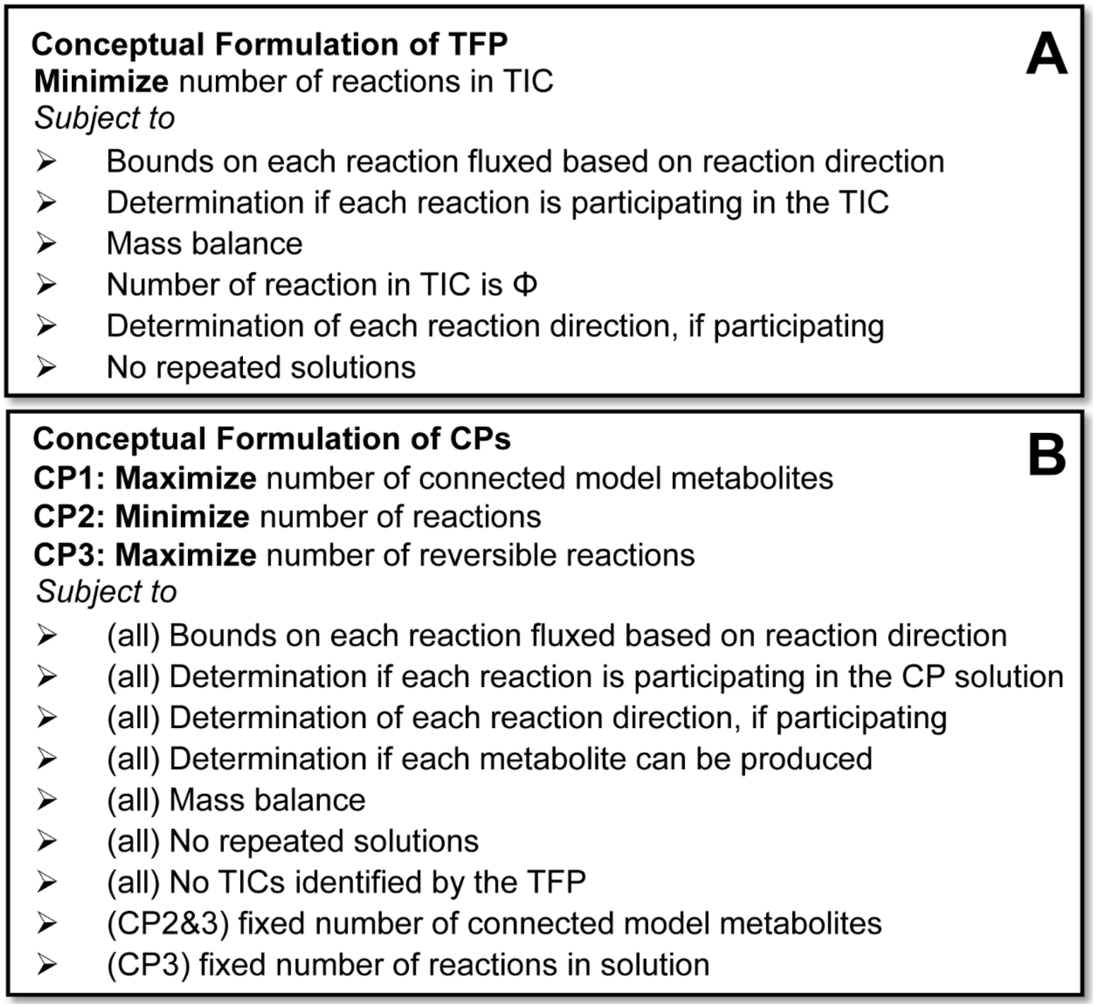


Figure 2.2: Workflow diagram of OptFill.

Extended Caption: This is a workflow diagram of the OptFill tool. Green nodes represent the preparatory workflow, blue the workflow of the TIC-Finding Problem (TFP), brown the workflow of the Connecting Problems (CPs), purple the error-handling workflow imbedded in the CPs workflow, and red the endpoints of the workflow. This color scheme is consistent with Figure 2.1. It should be noted that only one endpoint truly exists, when a solution to CP1 is not found, because the other problems, CP2 and CP3 will have solutions if CP1 has a solution, hence the workflow exit points being represented by a question mark at these points.



Conceptual Formulation of TFP

Minimize number of reactions in TIC

Subject to

- Bounds on each reaction fluxed based on reaction direction
- Determination if each reaction is participating in the TIC
- Mass balance
- Number of reaction in TIC is Φ
- Determination of each reaction direction, if participating
- No repeated solutions

Conceptual Formulation of CPs

CP1: Maximize number of connected model metabolites

CP2: Minimize number of reactions

CP3: Maximize number of reversible reactions

Subject to

- (all) Bounds on each reaction fluxed based on reaction direction
- (all) Determination if each reaction is participating in the CP solution
- (all) Determination of each reaction direction, if participating
- (all) Determination if each metabolite can be produced
- (all) Mass balance
- (all) No repeated solutions
- (all) No TICs identified by the TFP
- (CP2&3) fixed number of connected model metabolites
- (CP3) fixed number of reactions in solution

Figure 2.3: Conceptual formulation of OptFill.

Extended Caption: This figure give a conceptual formulation of the TIC-Finding Problem, TFP, in part (A) and the Connecting Problems, CP1, CP2, and CP3, in part (B). In part (B), as three connecting problems are solved, each conceptual constraint has indicated CPs to which it is applied. Conceptual constraints may require multiple mathematical constraints to be realized, see Transparent Method for mathematical formulation.

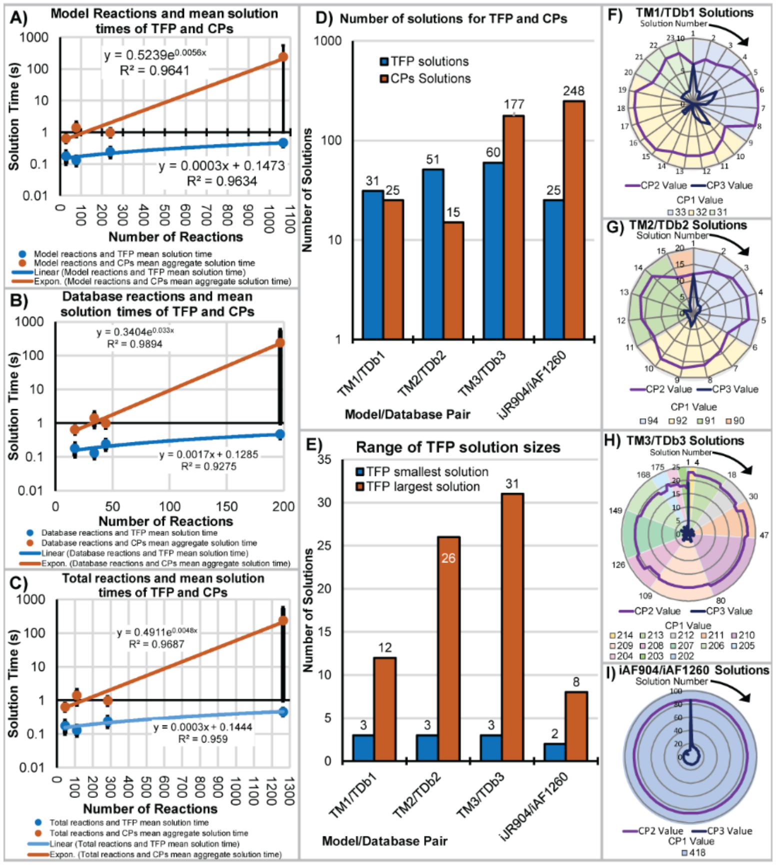


Figure 2.4: OptFill speed and solutions from first version of OptFill.

Extended Caption: This figure show the trends in solution time, (A) through (C), of the TIC-Finding Problem (TFP, blue) and the Connecting Problems (CPs, brown) with trend lines with the highest Pearson's correlation coefficient of linear, exponential, power, and logarithmic fits. These trends are considered with respect to the number of reactions in the model (A), database (B), and total reactions (C). Parts (D) through (F) highlights the trends of solutions. Part (D) highlights the number of solutions found by the TFP and CPs; Part (E) highlights the range in size of the identified potential TICs by the TFP. Parts (F) through (I) highlight the variety of CPs' solutions. In these figures, the pie-chart indicates the number of metabolites connected by the CP1 solution, and the radar chart is used to indicate the CP2 solution (number of reactions added) and the CP3 solution (number of those reactions that are added reversibly).

2.6. MATERIALS AND METHODS

2.6.1. Model-Database TIC-Finding Problem (TFP)

The first step of the OptFill method requires the iterative solving of the Mixed Integer Linear Programming (MILP) TIC-Finding Problem (TFP) applied to the model and database. This problem is defined below and is designed such that a TIC which could exist between the model and database with given reaction flux bounds will be a solution to the TFP.

$$\boldsymbol{minimize \sum_{j \in J} \eta_j} \tag{2.1}$$

Subject to (s.t.)

$$\boldsymbol{\beta_j v_j^{LB} + \epsilon \eta_j \leq v_j \leq (1 - \beta_j) v_j^{UB} - \epsilon \beta_j} \qquad \boldsymbol{\forall j \in J} \tag{2.2}$$

$$\boldsymbol{\eta_j v_j^{LB} \leq v_j \leq \eta_j v_j^{UB}} \qquad \boldsymbol{\forall j \in J} \tag{2.3}$$

$$\boldsymbol{\sum_{j \in J} S_{ij} v_j = 0} \qquad \boldsymbol{\forall i \in I} \tag{2.4}$$

$$\boldsymbol{\sum_{j \in J} \eta_j = \phi} \tag{2.5}$$

$$\boldsymbol{\sum_{j_{db} \in J^{DB}} \eta_{j_{db}} \geq 1} \tag{2.6}$$

$$\boldsymbol{\alpha_j \leq \eta_j} \qquad \boldsymbol{\forall j \in J} \tag{2.7}$$

$$\boldsymbol{\alpha_j \leq 1 - \beta_j} \qquad \boldsymbol{\forall j \in J} \tag{2.8}$$

$$\boldsymbol{\alpha_j \geq \eta_j - \beta_j} \qquad \boldsymbol{\forall j \in J} \tag{2.9}$$

$$\boldsymbol{\sum_{j \in J} \alpha_j \left(\alpha'_{s_f,j} \right) \leq \sum_{j \in J} \alpha'_{s_f,j} - \gamma_{s_f}} \qquad \boldsymbol{\forall s_f \in S_f} \tag{2.10}$$

$$\sum_{j\in J} \beta_j \left(\beta'_{s_f,j}\right) \leq \sum_{j\in J} \beta'_{s_f,j} - \left(1 - \gamma_{s_f}\right) \qquad \forall s_f \in S_f \qquad (2.11)$$

Fixed Values

$\epsilon = 1E - 3 \equiv a\ small\ number$

$v_j^{LB} \in \mathbb{R} \equiv lower\ bound\ of\ reaction\ j\ flux$

$v_j^{UB} \in \mathbb{R} \equiv upper\ bound\ of\ reaction\ j\ flux$

$S_{ij} \in \mathbb{R} \equiv stoichiometric\ coefficient\ of\ metabolite\ i\ in\ reaction\ j$

$$\alpha'_{s_f,j} = \begin{cases} 1\ if\ reaction\ j\ participates\ in\ previous\ TFP\ solution\ s_f\ with\ positive\ flux \\ 0\ otherwise \end{cases}$$

$$\alpha'_{s_f,j} = \begin{cases} 1\ if\ reaction\ j\ participates\ in\ previous\ TFP\ solution\ s_f\ with\ negative\ flux \\ 0\ otherwise \end{cases}$$

Variables

$$v_j \in \mathbb{R} \equiv flux\ of\ reaction\ j\ in \frac{mmol}{gDW \cdot h}$$

$$\eta_j = \begin{cases} 1\ if\ reaction\ j\ participates\ in\ current\ TIC\ solution \\ 0\ otherwise \end{cases}$$

$$\alpha_j = \begin{cases} 1\ if\ reaction\ j\ participates\ in\ current\ TIC\ solution\ with\ positive\ flux \\ 0\ otherwise \end{cases}$$

$$\beta_j = \begin{cases} 1\ if\ reaction\ j\ participates\ in\ current\ TIC\ solution\ with\ negative\ flux \\ 0\ otherwise \end{cases}$$

$\gamma_j \in [0, 1] \equiv binary\ value\ which\ ensures\ that\ at\ least\ 1\ integer\ cut\ holds$

The set s_f is the set of all previously found TICs and represents the solution space that is known. It should be noted that set J is the set of all reactions in the database and model, of which set J^{DB}, the set of all reactions in the database, is a subset. Further, it should be noted that I is the set of all metabolites in the database and the model. Parameters (fixed values) and variables are defined after all constraints have been listed. The TIC-finding problem is run with all nutrient uptakes turned off, so that any reaction flux is unrealistic and due to one or more TICs. The TFP is included in File S3 as GAMS (Generalized Algebraic Modeling System) code. The following subsections will describe the above equations constituting the TFP in detail.

2.6.1.1. Objective function and sought TIC size

The solution of the TFP is itself a TIC. The objective function, equation (2.1), is minimization of the number of reactions participating in the TIC solution. This objective function is irrelevant in the solution due to equation (2.5), as equation (2.5) specifies the size of the TIC sought, and thus the objective function value, and is included to ensure that each possible TIC size is investigated. The order of solutions, when the workflow in Figure 2.2 is followed, is unimportant, and may vary each time the TFP is applied to a model.

2.6.1.2. Enforcing flux bounds and reaction participation

Equations (2.2) and (2.3) are constraints which enforce the given reaction flux bounds and determine if a reaction participates in the identified TIC. The variable η_j stores if a reaction participates in a TIC, while variables α_j and β_j store direction of participation. Reaction flux bounds v_j^{LB} and v_j^{UB} are determined manually based on reaction direction (reversible, irreversible forward, or irreversible backward), limitations on nutrient uptake rates, and reaction state (either on or off depending on genotype, nutrient availability). Equation (2.6) ensures that at least one database reaction holds flux. Equation (2.2) specifically identifies if reaction j participates in the solution TIC by requiring some small, minimum reaction flux, ϵ, for participating reactions such that equation (2.12) is true. Further, it identifies the direction of that reaction.

$$|\boldsymbol{v_j}| \geq \boldsymbol{\epsilon} \qquad \forall \boldsymbol{j} \in \{\boldsymbol{J}|\boldsymbol{\eta_j} = \mathbf{1}\} \tag{2.12}$$

Equation (2.3) ensures that if any reaction does not meet the reaction flux threshold to participate in the TIC solution, that the reaction flux is constraint to zero.

2.6.1.3. Identifying positive flux participation in the TIC

Equations (2.7) through (2.9) are a linearized version of the following statement.

$$\boldsymbol{\alpha_j} \leq \boldsymbol{\eta_j}(\mathbf{1} - \boldsymbol{\beta_j}) \tag{2.13}$$

The linearization in equations (2.7) through (2.9) functions the same as (2.13) because η_j and β_j are binary variables. This linearization is made in order to preserve the linear nature of the TFP. A linear optimization problem can guarantee both global solution optimality and that all solutions in the solution space can be enumerated, which in this case guarantees that all TICs are found of a given size.

2.6.1.4. Integer Cuts for Repeated Solutions

Equations (2.10) and (2.11) are integer cuts which prevent repetition of solutions. It should be noted that these repeated solutions include direction. Therefore, to be identified as the same TIC, the set of participating reactions and the directions in which they participate must be the same. Consider the following set of chemical equations for an illustration of how these integer cuts prevent repeated solutions.

$1\,A \leftrightarrow 1\,B$

$1\,B \leftrightarrow 1\,C$

$1\,C \leftrightarrow 1\,A$

The TFP, because of these integer cuts, would identify two TICs existing in this set of chemical reactions. The first would be all reactions listed above proceeding in the forward direction, while the second would be all reactions listed above proceeding in the backward direction. These are identified separately because their reaction directions are different, although the participating reactions are the same.

2.6.2. Modified TIC-Finding Problem (mTFP)

The TFP can be modified for the identification of TICs inherent to a metabolic model to aid in model curation. The modified TIC-Finding Problem (mTFP) can be formulated via equations (2.1) through (2.5) and equations (2.7) through (2.11). All set, parameter, and variable definitions are the same as in the unmodified TFP.

2.6.3. First Connecting Problem (CP1)

The connecting problems are the series of optimization problems which are solved following the solving of the TFP. First discussed will be the first Connecting Problem (CP1). The solution to a CP is a set of database reactions which, when added to the model, will increase model connectivity. The solution to CP1 gives the maximum number of model metabolites which could be connected using the database. The formulation of CP1 is given below.

$$\boldsymbol{maximize\ Z_{met}} = \sum_{i_m \in I^M} \boldsymbol{x_{i_m}} \tag{2.14}$$

Subject to

$$\sum_{j_{db} \in J^{DB}} \boldsymbol{\zeta_{j_{db}} \geq 1} \tag{2.15}$$

$$\boldsymbol{\rho_{j_{db}} v_{j_{db}}^{LB} \leq v_{j_{db}} \leq \delta_{j_{db}} v_{j_{db}}^{UB}} \qquad \forall \boldsymbol{j_{db}} \in \boldsymbol{J^{DB}} \tag{2.16}$$

$$\boldsymbol{\theta_j v_j^{LB} \leq v_j \leq (1-\theta_j) v_j^{UB} - \epsilon\theta_j} \qquad \forall \boldsymbol{j} \in \boldsymbol{J} \tag{2.17}$$

$$\boldsymbol{(1-\lambda_j) v_j^{LB} + \epsilon\lambda_j \leq v_j \leq \lambda_j v_j^{UB}} \tag{2.18}$$

$$\boldsymbol{x_i} \leq \sum_{j \in J} \boldsymbol{[\lambda_j \xi_{i,i} + \theta_j \psi_{i,j}]} \qquad \forall \boldsymbol{i} \in \boldsymbol{I} \tag{2.19}$$

$$\boldsymbol{x_b = 1} \qquad \forall \boldsymbol{b} \in \boldsymbol{B} \subset \boldsymbol{I} \tag{2.20}$$

$$\sum_{j \in J} \boldsymbol{S_{ij} v_j = 0} \qquad \forall \boldsymbol{i} \in \boldsymbol{I} \tag{2.21}$$

$$\boldsymbol{\zeta_{j_{db}}} \leq \sum_{i \in I} \boldsymbol{[\lambda_j \xi_{i,i} + \theta_j \psi_{i,j}]} \qquad \forall \boldsymbol{j_{db}} \in \boldsymbol{J^{DB}} \tag{2.22}$$

$$\boldsymbol{\delta_{j_{db}} + \rho_{j_{db}} - \omega_{j_{db}} = \zeta_{j_{db}}} \qquad \forall \boldsymbol{j_{db}} \in \boldsymbol{J^{DB}} \tag{2.23}$$

$$\boldsymbol{\omega_{j_{db}} \leq \delta_{j_{db}}} \qquad \forall \boldsymbol{j_{db}} \in \boldsymbol{J^{DB}} \tag{2.24}$$

$$\boldsymbol{\omega_{j_{db}} \leq \rho_{j_{db}}} \qquad \forall \boldsymbol{j_{db}} \in \boldsymbol{J^{DB}} \tag{2.25}$$

$$\sum_{j_{db}\in J^{DB}} \boldsymbol{\delta}_{j_{db}}\left(\boldsymbol{\delta}'_{s_c,j_{db}}\right) \leq \sum_{j_{db}\in J^{DB}} \boldsymbol{\delta}'_{s_c,j_{db}} - \boldsymbol{\sigma}_{s_c} \qquad \forall \boldsymbol{s_c} \in \boldsymbol{S_c} \qquad (2.26)$$

$$\sum_{j_{db}\in J^{DB}} \boldsymbol{\rho}_{j_{db}}\left(\boldsymbol{\rho}'_{s_c,j_{db}}\right) \leq \sum_{j_{db}\in J^{DB}} \boldsymbol{\rho}'_{s_c,j_{db}} - \left(\mathbf{1} - \boldsymbol{\sigma}_{s_c}\right) \qquad \forall \boldsymbol{s_c} \in \boldsymbol{S_c} \qquad (2.27)$$

$$\sum_{j_{db}\in J^{DB}} \left(\boldsymbol{\delta}'_{s_c,j_{db}} - \boldsymbol{\delta}_{j_{db}}\right) + \sum_{j_{db}\in J^{DB}} \left(\boldsymbol{\rho}'_{s_c,j_{db}} - \boldsymbol{\rho}_{j_{db}}\right) \geq \left(\sum_{j_{db}\in J^{DB}} \boldsymbol{\omega}'_{s_c,j_{db}}\right) + \mathbf{1} \qquad \forall \boldsymbol{s_c} \in \boldsymbol{S_c} \qquad (2.28)$$

$$\sum_{j_{db}\in J^{DB}} \boldsymbol{\delta}_{j_{db}}\left(\boldsymbol{\alpha}'_{s_f,j_{db}}\right) \leq \sum_{j_{db}\in J^{DB}} \boldsymbol{\alpha}'_{s_f,j_{db}} - \boldsymbol{\tau}_{s_f} \qquad \forall \boldsymbol{s_f} \in \boldsymbol{S_f} \qquad (2.29)$$

$$\sum_{j_{db}\in J^{DB}} \boldsymbol{\rho}_{j_{db}}\left(\boldsymbol{\beta}'_{s_f,j_{db}}\right) \leq \sum_{j_{db}\in J^{DB}} \boldsymbol{\beta}'_{s_f,j} - \left(\mathbf{1} - \boldsymbol{\tau}_{s_f}\right) \qquad \forall \boldsymbol{s_f} \in \boldsymbol{S_f} \qquad (2.30)$$

$\boldsymbol{Fixed\ Values\ Unique\ to\ CP}\mathbf{1}$

$\boldsymbol{M} = \mathbf{1}\boldsymbol{E}\mathbf{3} \equiv \boldsymbol{a\ very\ large\ number}$

$$\boldsymbol{\delta}'_{s_c,j_{db}} = \begin{cases} \mathbf{1}\ \boldsymbol{if\ reaction\ j_{db}\ is\ added\ in\ the\ forward\ direction\ from\ the} \\ \boldsymbol{database\ in\ solution\ s_c} \\ \mathbf{0}\ \boldsymbol{otherwise} \end{cases}$$

$$\boldsymbol{\rho}'_{s_c,j_{db}} = \begin{cases} \mathbf{1}\ \boldsymbol{if\ reaction\ j_{db}\ is\ added\ in\ the\ backward\ direction\ from\ the} \\ \boldsymbol{database\ in\ solution\ s_c} \\ \mathbf{0}\ \boldsymbol{otherwise} \end{cases}$$

$$\boldsymbol{\omega}'_{s_o,j_{db}} = \begin{cases} \mathbf{1}\ \boldsymbol{if\ reaction\ j_{db}\ is\ added\ in\ the\ forward\ direction\ from\ the} \\ \boldsymbol{database\ in\ solution\ s_o} \\ \mathbf{0}\ \boldsymbol{otherwise} \end{cases}$$

$$\boldsymbol{\xi}_{i,j} = \begin{cases} \mathbf{1}\ \boldsymbol{if\ metabolite\ i\ is\ on\ the\ RHS\ of\ reaction\ j}\ \left(\boldsymbol{S}_{i,j} > \mathbf{0}\right) \\ \mathbf{0}\ \boldsymbol{otherwise} \end{cases}$$

$$\boldsymbol{\psi}_{i,j} = \begin{cases} \mathbf{1}\ \boldsymbol{if\ metabolite\ i\ is\ on\ the\ LHS\ of\ reaction\ j}\ \left(\boldsymbol{S}_{i,j} < \mathbf{0}\right) \\ \mathbf{0}\ \boldsymbol{otherwise} \end{cases}$$

$$Variables\ Unique\ to\ CP1$$

$$\delta_{j_{db}} = \begin{cases} 1\ if\ reaction\ j_{db}\ is\ added\ in\ the\ forward\ direction\ from\ the\ database \\ 0\ otherwise \end{cases}$$

$$\rho_{j_{db}} = \begin{cases} 1\ if\ reaction\ j_{db}\ is\ added\ in\ the\ backwards\ direction\ from\ the\ database \\ 0\ otherwise \end{cases}$$

$$\omega_{j_{db}} = \begin{cases} 1\ if\ reaction\ j_{db}\ is\ added\ reversibly\ from\ the\ database\ (\delta_{j_{db}} = \rho_{j_{db}} = 1) \\ 0\ otherwise \end{cases}$$

$$\zeta_{j_{db}} = \begin{cases} 1\ if\ reaction\ j_{db}\ is\ part\ of\ the\ solution \\ 0\ otherwise \end{cases}$$

$$\theta_j = \begin{cases} 1\ if\ reaction\ is\ proceding\ in\ backwards\ direction\ \left(v_j < 0\right) \\ 0\ otherwise \end{cases}$$

$$\lambda_j = \begin{cases} 1\ if\ reaction\ is\ proceding\ in\ forwards\ direction\ \left(v_j > 0\right) \\ 0\ otherwise \end{cases}$$

$$x_i = \begin{cases} 1\ if\ connected\ model\ produces\ metabolite\ i \\ 0\ otherwise \end{cases}$$

$$\sigma_{s_c} \in [0,1] \equiv binary\ variable\ which\ ensures\ that\ the\ solution\ is\ unique\ from\ previous\ solutions\ in\ at\ least\ one\ direction\ of\ one\ database\ reaction$$

$$\tau_{s_c} \in [0,1] \equiv binary\ variable\ which\ ensures\ that\ the\ solution\ is\ free\ from\ TICs\ in\ that\ at\ least\ one\ direction\ of\ one\ database\ reactions\ which\ could\ cause\ a\ TIC\ is\ added\ in\ the\ TIC - causing\ direction\ for\ each\ TIC\ identified\ by\ the\ TFP$$

Where I^M is defined as the set of metabolites in the model and is a subset of I. When CP1 is solved, the optimal value of Z_{met} is the maximum number of metabolites which can be connected in the model by adding reactions from the database, given all previous solutions (if any) and all identified potential TICs. It should be noted that all sets and parameters have the same definitions here as in the TFP, with the additions of J^M being the set of model reactions which is a subset of J, of I^M being the set of model metabolites which is a subset of I, s_c being the set of all previous connecting problem solutions, s_o being the set of all previous connecting problem solutions with at least one reversible reaction being added from the database, and B being the set of all metabolites which are involved in the biomass equation which is a subset of I.

The following statements give, broadly, the rational for each constraint equation. Equation (2.14) ensures that at least one reaction is added from the database for each solution. Equation (2.15) ensures that each database reaction only has flux if it is added. Equation (2.16) ensures that the user-defined reaction flux bounds hold. Equations (2.17) through (2.19) determine which metabolites the fixed model can produce, equation (2.19) ensures that the fixed model can produce biomass. Equation (2.20) ensures mass balance. Equation (2.21) ensures that added reactions are productive, e.g. that the added reaction does produce one or more metabolites. Equations (2.22) through (2.24) ensure that each database reaction for the connecting solution is added as a forward, backward, or reversible reaction (e.g. both as a forward and a backward reaction). Equations (2.25) though (2.28) are integer cuts preventing repeated solutions, while Equations (2.29) and (2.30) are integer cuts preventing the full addition of a TIC through the CP solution. The following subsections will describe some of the above equations constituting the CPs in greater detail. The CPs are included in File S4 as GAMS (Generalized Algebraic Modeling System) code. The following subsections will describe some of the above equations constituting the CPs in greater detail.

a

2.6.3.1. Determination of Metabolite Production

Important to CPs is the determination of whether or not a metabolite is produced in the connected model. Equations (2.17) and (2.18) are used to determine which direction reactions proceed in the connected model. Equation (2.19) essentially states that a metabolite is produced if at least one reaction produces that metabolite by having flux in the direction of that metabolite (either through backwards flux and a negative stoichiometric coefficient or forward flux and a positive stoichiometric coefficient). Equation (2.20) ensures that all metabolites necessary for growth (those involved in biomass production) are produced, as all models of metabolism should be capable of producing biomass, even if biomass is not ultimately the objective used. For instance, alternate objectives could include the maximization of production of a given metabolite (Herrgård, Fong, & Palsson, 2006)(Price, Reed, & Palsson, 2004), the minimization of the uptake of a particular substrate (Gomes de Oliveira Dal'Molin et al., 2015), or minimization of metabolic adjustment (MOMA) (Herrgård et al., 2006)(Price et al., 2004). Ultimately, each objective type some fixed or variable non-zero level of biomass production and therefore all models require some ability to grow, making these constraints reasonable for reconstructions regardless of the ultimate objective used. Equation (2.22) ensures that reactions added from the database are productive, e.g. that each added reaction is capable of producing at least one metabolite. This constraint ensures that reactions incapable of carrying flux are not added to the model.

2.6.3.2. Direction of Added Database Reactions

Equations (2.22) through (2.25) largely deal with the direction in which reactions are added from the database. Equations (2.22) ensures that reactions added from the database are productive. Equation (2.23) ensures that ζ_{jdb} is equal to 1 if reaction j_{db} is added to the model as part of this solution, and zero otherwise. Equations (2.23) through (2.25) are the linearization of the multiplication of two binary variables stated below.

$$\boldsymbol{\omega_{j_{db}} = \delta_{j_{db}}\rho_{j_{db}}} \tag{2.31}$$

This linearization is done for the same reasons that the TFP has been linearized. The sum of these constraints ensures that any reaction added reversibly is treated as a reaction added both forward and backwards for the purposes of integer cuts to avoid repeated solutions.

2.6.3.3. Integer Cuts for Repeated Solutions

Equations (2.26) through (2.28) define integer cuts used to avoid repeat solutions. Equations (2.26) and (2.27) have been designed on similar lines to (2.10) and (2.11), designed to avoid repeat solutions. Through the integer cuts in equations (2.26) through (2.28), both the reactions and their directions are integral to the solution; therefore, any different between solutions in reaction direction or reactions included is recorded as a second solution. Equation (2.28) prevents the repetition of a solution that could be caused by changing a reversible database reaction addition into an irreversible one.

2.6.3.4. Integer Cuts for TIC-less Connecting

Equations (2.29) through (2.30) define integer cuts which ensure that a TIC is not added to the connecting solution. This is done by considering both reaction identity and direction for both the addition of database reactions and for the avoidance of TICs. This results in a minimum perturbation to the solution space of CPs caused by each TIC. As with other directional integer cuts, only one cut needs be in effect at minimum in order to define a new solution.

2.6.3.5. Modified First Connecting Problem

A modified CP1 was used to get an initial count of the maximum number of metabolites which the raw model can produce. This modified CP1 made use of equations (2.14), and (2.16) through (2.30). In place of equation (2.15) the following equation was used to ensure that no database reactions were considered in maximizing the number of metabolites which may be connected.

$$\sum_{j_{db}\in J^{DB}} \zeta_{j_{db}} = \mathbf{0} \tag{2.32}$$

2.6.4. Second Connecting Problem

The second Connecting Problem (CP2) is defined as equations (2.15) through (2.30) with the addition of the objective function and constraint equation (2.34) stated below.

$$\boldsymbol{minimize}\ \boldsymbol{Z_{rxn}} = \sum_{j_{db}\in J^{DB}} \zeta_{j_{db}} \tag{2.33}$$

s.t.

Equations (2.15) through (2.30)

$$\sum_{i_m\in I^M} x_{i_m} = Z_{met,opt} \tag{2.34}$$

Where $Z_{met,opt}$ is defined as the optimal objective value of CP1. When CP2 is solved, the optimal value of Z_{rxn} is the minimum number of reactions which, when added from the database, can connect the previously determined maximum number of model metabolites, given all previous solutions (if any) and all identified potential TICs.

2.6.5. Third Connecting Problem

The third Connecting Problem (CP3) is defined as equations (2.15) through (2.30), equation (2.34), and constraint equation (2.36) stated below.

$$\boldsymbol{maximize\ Z_{rev}} = \sum_{\boldsymbol{j_{db} \in J^{DB}}} \boldsymbol{\omega_{j_{db}}} \tag{2.35}$$

s.t.

Equations (2.15) through (2.31), (2.34)

$$\sum_{\boldsymbol{j_{db} \in J^{DB}}} \boldsymbol{\zeta_{j_{db}}} = \boldsymbol{Z_{rxn,opt}} \tag{2.36}$$

Where $Z_{rxn,opt}$ is defined as the optimal objective value of CP2. When CP3 is solved, the optimal value of Z_{rev} is the maximum number of reversible reactions which can be used to achieve the minimum number of reaction additions to maximize model connectivity, given all previous solutions (if any) and all identified potential TICs. The solution of CP3 is the solution accepted as optimal.

CP3 has been found to be needed due to allowing database reactions to be added forward, backward, and reversibly. Since adding a reaction reversibly rather than irreversibly in some cases has made no difference, this resulted in an inconsistent number of solutions to the set of CPs. Therefore, in one run two solutions would be returned (the irreversible solution has been returned, then the reversible), where in a subsequent run perhaps only one solution would be returned if the reversible solution has been returned first, and then integer cuts (2.26) and (2.27) would preclude the irreversible solution. This third connecting problem has been added to deal with such situations by forcing the reversible solution to be returned first, resulting in a standardized, minimized set of solutions.

2.6.6. FBA of Connected Model

Once the CPs have been solved and the identity and direction of models to be added from the database to the model for a given solution are known, Flux Balance Analysis (FBA) is performed on the connected model. As the models are not physically merged, this takes the following form.

$$\boldsymbol{maximize\ Z_{bio} = v_{biomass}} \tag{2.37}$$

s.t.

$$\sum_{j \in J} \boldsymbol{S_{ij} v_j = 0} \qquad \forall \boldsymbol{i} \in \boldsymbol{I} \tag{2.38}$$

$$\boldsymbol{v_{j_m}^{LB} \le v_{j_m} \le v_{j_m}^{UB}} \qquad \forall \boldsymbol{j_m} \in \boldsymbol{J^M} \tag{2.39}$$

$$\boldsymbol{\rho'_{j_{db}, s_{curr}} v_{j_{db}}^{LB} \le v_{j_{db}} \le \delta'_{j_{db}, s_{curr}} v_{j_{db}}^{UB}} \qquad \forall \boldsymbol{j_{db}} \in \boldsymbol{J^{DB}} \tag{2.40}$$

All variables, parameters, and sets are the same as in previous equations, and in addition s_{curr} represents the current connecting solution. In the above formulation, equation (2.39) takes into account the current solution of the CPs. A biomass maximization objective function was chosen for this work, but other objective could be selected depending on what part of metabolism is of most interest.

2.6.7. Creation of Test Models and Databases

Test model have been created in tandem with their databases using KEGG maps of pathways to identify sets of reactions which might produce a functional metabolic model. The first Test Model (TM1) and Test Database (TDb1) have been built from the “starch and sucrose metabolism” (map00500) and the “glycolysis/gluconeogenesis” (map00010) metabolic maps with the goal of producing a minimal prokaryotic model which utilizes sucrose, produces ethanol and biomass, and has some TICs which exist between the database and model where TM1 cannot produce biomass (without some TDb1 reactions) and contains no inherent TICs. Since only sucrose metabolism and glycolysis have been included in this model, biomass for this model is based on glucose, fructose, and an arbitrary growth-associated maintenance (GAM) value of 2. The coefficient of glucose in the biomass equation has then been scaled such that the molecular weight of biomass is 1000 g/mol. Non-Growth Associated Maintenance (NGAM) has also been defined arbitrarily as 2. TM1 and TDb1 have been constructed rationally with as many reversible reactions as possible, such that 22 of the 28 reactions are reversible in TM1 and all 17 reactions are reversible in TDb1. Once TM1 and TDb1 have been constructed, OptFill has been applied to them. This has resulted in the identification of 31 TICs consisting of 3 to 12 reactions by the TFP using the CPLEX solver. See results section for detail.

The first solution reported by OptFill for TM1/TDb1 has been added to TM1 to create the initial second Test Model (TM2). Added manually to this initial TM2 model is portions of the “pentose phosphate” pathway (map00030) and fatty acid biosynthesis” (map00061) pathway. The biomass equation has been updated to include a small amount (stoichiometric coefficient 0.01) of fatty acid products (8-, 10-, 12-, 14-, 16-, and 18-carbon fatty acid products) and the coefficient of glucose has again been adjusted to ensure biomass molecular weight was 1000 g/mol. Certain reactions in both pathways have been selected to constitute the second Test Database (TDb2), again with the aim of being a small prokaryotic model which utilizes sucrose, produces ethanol, produces biomass, and has some TICs which exist between the database and model where TM2 cannot produce biomass (without some TDb2 reactions) and contains no inherent TICs. In total, TM2 consists of 77 reactions (with 65 being reversible), and TDd2 consists of 34 reactions (all reversible). Once TM2 and TDb2 have been constructed, OptFill has been applied to them, see results section for details.

As with the construction of TM2, the third Test Model (TM3) has initially been constructed from the first solution of OptFill applied to TM2/TDb2 added to a test model. This test model has then been expanded to include "nitrogen metabolism" (map00910, with ammonium uptake), "sulfur metabolism" (map00920, with sulfate uptake), and synthesis pathways for all 20 amino acids. The biomass equation has been updated to include a small amount (stoichiometric coefficient 0.1) of each of the 20 primary amino acids, following which the coefficient of glucose has again been adjusted to ensure biomass molecular weight was 1000 g/mol. Unlike previous test models, this working test model (e.g. capable of producing biomass) with some TICs has first been developed, split between reactions belonging to TM2 or OptFill solution thereof, and "other" reactions. Then each of these "other" reaction has been assigned a random value (between 0 and 1) and those with a value greater than or equal to 0.7 have been assigned to the third Test Database (TDb3), and those with a value less than or equal to 0.8 have been assigned to the third Test Model (TM3). The code to perform this is included as part of the GitHub OptFill (10.52.81/zenodo.8475) or Mendeley Data (10.17632/npdwbmb7d7.1) repositories accompanying this work. Following this, the mTFP has been applied to TM3 in order to ensure that the model is TIC-less. For removing TICs from TM3, the number of occurrences of each reaction participating in all TICs has been counted, that has the highest occurrence, excluding those reactions from TM2 and TDb2, has been moved to TDb3. In the case of ties, the reaction with the highest reaction ID number has been moved to TDb3. In total, TM3 consists of 210 reactions (196 reversible), and TDb3 consists of 77 reactions (all reversible). Once TM3 and TDb3 have been constructed, OptFill has been applied to them, see results section for details.

It should be noted that for all instances of OptFill applied to test models some low number of execution errors have been allowed, five are allowed in this example option allowing execution errors: “execerr=5”. This has been done because GAMS throws an execution error if the RHS and LHS of a constraint are fixed and those fixed values do not satisfy the constraint. In the case of OptFill, this is not necessarily an issue, as it simply indicates that there are no more feasible solutions and that the program should continue onto the next problem or step. Graphical summaries comparing project runtimes have then been generated in Table S2. Result summaries, graphs and biomass calculations related to Figure 2.4 (Microsoft Excel) to produce Figure 2.4. Trend line and Pearson correlation values included in this figure have been generated automatically by Microsoft Excel. Linear, logarithmic, exponential, and power trend lines have been investigated, and the best fit line is displayed for each dataset. Polynomial trend lines have not been investigated as these trend lines can lead to overfitting errors.

2.6.8. Application of *i*AF1260 to *i*JR904

In the application of OptFill to published *Escherichia coli* GSMs, *i*JR904 (Reed et al., 2003) was treated as the model and *i*AF1260 (Feist et al., 2007) as the source of reactions to build the database for OptFill. Minor formatting of both of these models was accomplished using the code in the GitHub OptFill or Mendeley Data repositories accompanying this work. Such formatting changes include changing of how reaction arrows appeared and location of metabolite compartment notation. Following this formatting, all exchange reactions were removed from *i*AF1260, as it was decided to use the media definition provided for *i*JR904 by Reed *et al.*, 2003, specifically for the case of aerobic growth on acetate. Whereas very large bounds in *i*JR904 have been defined as $1e^{30}$, these have been redefined as $1e^{3}$ as both quantities are sufficiently large in the context of GSMs to be a red flag should any reaction flux reach that quantity. Further, $1e^{3}$ is the value of *M* used elsewhere in the code, resulting in a standard value for a “very large number”.

Once the aforementioned changes had been made, *i*AF1260 (sans exchange reactions) and *i*JR904 were compared in Table the GitHub OptFill or Mendeley Data repositories accompanying this work so that reactions that are in both model would be removed from *i*AF1260. These modifications resulted in 1441 reactions remaining in the initial *i*AF1260-based database. The initial iAF1260-based database is provided in the GitHub OptFill or Mendeley Data repositories accompanying this work, as is the GAMS code used in this application of OptFill. The OptFilling of *i*JR904 using an *i*AF1260-based database is different from the code used for the test models/database only in formatting of the output file (identifiers used were considerably longer than KEGG identifiers causing formatting issues). This was allowed for seven days to attempt to solve, in which time it did not return a single CPs solution; therefore, it was decided that the database needed to be made smaller. Both the initial *i*AF1260-based database and *i*JR904 were combined into a single pseudo-model file, to which Flux Variability Analysis (FVA) was applied. Those reactions which hold flux, 715 reactions, formed the second iAF1260-based database.

OptFill was applied to this second database, but still resulted in very long solution times; therefore, those reactions which participated in solutions which were achieved in 24 hours (four solutions) were chosen to form the final *i*AF1260-based database. This database consists of 182 reactions. It was found that this resulted in no TFP solutions; therefore, six more reactions were added to produce a database which had 25 potential TICs with the *i*JR904 model. OptFill was then applied to *i*JR904 using this final *i*AF1260-based database of 188 reactions.

2.6.9. CPLEX Solver Options

As the order of solutions presented is important, solver options which allowed non-optimal solutions or created relaxations by which the truly optimal solution could not be reached, or a suboptimal solution would be accepted, were disabled. In particular, the infeasibility gap was set to the lowest possible value, small infeasibilities were disallowed, no relaxation was allowed in the value of integers, no optimality gap was allowed in the solution, and solver cuts which could result in non-optimal solutions were disabled. These cuts included zero-half, flow, clique, cover, mixed integer rounding, GUB cover, and Gomory fractional cuts. While the lack of these relaxation options and cuts no doubt increased solution time, these relaxations would decrease solution accuracy and order which was deemed unacceptable. The list of CPLEX relaxations used in this work can be found in the GitHub OptFill or Mendeley Data repositories accompanying this work.

Chapter 3

3. COMPUTATION-DRIVEN ANALYSIS OF MODEL POLYEXTREMOTOLERANT FUNGUS EXOPHIALA DERMATITIDIS: DEFENSIVE PIGMENT METABOLIC COSTS AND HUMAN APPLICATIONS

Portions of this material have previously appeared in the following publication:

W. L. Schroeder, S. D. Harris, and R. Saha, Computation-Driven Analysis of Model Polyextremotolerant Fungus Exophiala dermatitidis: Defensive Pigment Metabolic Costs and Human Applications, iScience, 23(2020) 1-17. Used with permission.

W. L. Schroeder, R. Saha, Protocol for Genome-Scale Reconstruction and Melanogenesis Analysis of Exophiala dermatitidis, STAR Protocols, 1(2020) 1-37. Used with permission.

3.1. PREFACE

The polyextremotolerant black yeast Exophiala dermatitidis is a tractable model system for investigation of adaptations that support growth under extreme conditions. Foremost among these adaptations are melanogenesis and carotenogenesis. A particularly important question is their metabolic production cost. However, investigation of this issue has been hindered by a relatively poor systems-level understanding of E. dermatitidis metabolism. To address this challenge, a genome-scale model (iEde2091) was developed. Using iEde2091, carotenoids were found to be more expensive to produce than melanins. Given their overlapping protective functions, this suggests that carotenoids have an underexplored yet important role in photo-protection. Furthermore, multiple defensive pigments with overlapping functions might allow E. dermatitidis to minimize cost. Because iEde2091 revealed that E. dermatitidis synthesizes the same melanins as humans and the active sites of the key tyrosinase enzyme are highly conserved this model may enable a broader understanding of melanin production across kingdoms.

3.2. INTRODUCTION

Extremophiles are organisms that can live in extreme conditions of temperature, acidity, alkalinity, or salinity. Studying these organisms not only expands our knowledge on the diversity of life but can also provide significant insights into how organisms adapt to stress, particularly metabolic and regulatory responses. *Exophiala dermatitidis* (hereafter, *Exophiala* or *E. dermatitidis*, also known as *Wangiella dermatitidis*), a highly-melanized black fungus and perhaps best known for its *H. sapiens* (hereafter, human) pathogenic properties (Paolo *et al.*, 2006; Poyntner *et al.*, 2016; Sudhadham *et al.*, 2008), is a potential model extremophile system due to its small genome of 26.4 Mb ("Exophiala dermatitidis NIH/UT8656 Genome Assembly," 2011) and its demonstrated extremotolerance with respect to temperature (heat and cold) (Paolo *et al.*, 2006; Sudhadham *et al.*, 2008), acidic pH (Sudhadham *et al.*, 2008; Chen *et al.*, 2014), light (Chen *et al.*, 2014; Nosanchuk and Casadevall, 2006a; Geis and Szaniszlo, 1984), and radiation (Chen *et al.*, 2014; Nosanchuk and Casadevall, 2006a; Geis and Szaniszlo, 1984), oxidative stress (Chen *et al.*, 2014; Geis and Szaniszlo, 1984) and likely tolerance to toxic heavy metals (Nosanchuk & Casadevall, 2006), harmful aromatic compounds (Moreno, Vicente, & de Hoog, 2018), various toxins (Moreno et al., 2018), antimicrobial compounds (Nosanchuk & Casadevall, 2006), and other stressors (nutrient, osmotic, and mechanical) (Moreno et al., 2018). The ability of *Exophiala* to adapt to most of these conditions seemingly results from two classes of defensive pigments: melanins, a class of pigments consisting of six-carbon ring monomers, and carotenoids, a class of polyisoprenoid pigments. *Exophiala* can produce three different types of melanin: i) 1,8-dihydroxynaphthalene melanin (hereafter, DHN-melanin), also called naphthalene melanin, ii) DOPA-melanin, also known as eumelanin (S. Ito & Wakamatsu, 2011) and iii) pyomelanin. Among these, DHN-melanin and pyomelanin are generally produced by fungi (Solano, 2014) including *Exophiala*, whereas eumelanin is produced by both fungi and animals, including humans (Ito and Wakamatsu, 2011; Solano, 2014). The combination of its small genome ("Exophiala dermatitidis NIH/UT8656 Genome Assembly," 2011), its ability to be cultured as yeast cells (Chen *et al.*, 2014; Ohkusu *et al.*, 1999), and production of eumelanin (S. Ito & Wakamatsu, 2011) makes *Exophiala* a potential model organism for human melanocytes, the cells in humans which produce melanins. Melanocytes are specialized cells in humans which are found primarily in the skin which produce pheomelanin and eumelanin in specialized subcellular organelles called melanosomes.

From outlined uses of GSMs in the first chapter, the reconstruction of a GSM of *Exophiala* can be a useful tool to investigate its potential as a model organism both for polyextremotolerant organism and for human melanocytes. However, GSMs are challenging to reconstruct for under-studied organisms such as *Exophiala*, where only approximately 43% of genes have some level of annotation (not including hypothetical or putative proteins), and less than 5% of genes are annotated with Enzyme Classification (EC) numbers which might be used to establish GPR links (*Exophiala dermatitidis NIH/UT8656 Genome Assembly*, 2011; *Exophiala dermatitidis (strain ATCC34100/CBS 525.76/NIH/UT8656)*, 2018). This lack of annotations often leaves large gaps in metabolic reconstructions which requires further scrutiny. One tool that we recently have developed is OptFill (Schroeder & Saha, 2020a), which performs whole-model Thermodynamically Infeasible Cycle (TIC) free gapfilling. TICs are detrimental to GSMs as they result in the reporting of unrealistic flux results, cause difficulties in using dual formulations of optimization problems (such as in this work), and can make energy costs such as ATP maintenance meaningless (Schroeder & Saha, 2020a). OptFill works by first identifying possible TICs which can occur between a database of functionalities proposed to fix the gaps in the model and the model itself. Then the reaction flux in the direction which would allow a TIC is excluded in the second step of OptFill, which attempts to maximize the number of model reactions fixed by adding new reactions (Schroeder & Saha, 2020a). Ultimately, this allows for the maximization of model connectivity while minimizing new functionalities added to the model, as well as opportunity to hypothesize functions for un- or poorly-annotated genes through the concurrent use of tools such as BLASTp (Altschul *et al.*, 1997; Altschul *et al.*, 2005). Through the process of reconstructing a GSM, metabolic pathways are thoroughly investigated, particularly those related to the subjects of the study, in this case defensive pigments. In addition, this reconstruction provides the basis for comparison between humans and *Exophiala*, which when supplemented with sequence alignment tools such as COBALT (Papadopoulos & Agarwala, 2007) can provide initial comparisons for determining the suitability of *E. dermatitidis* as a model organism.

Once a GSM is reconstructed, optimization-based tools of analysis may be applied to investigate *E. dermatitidis* as a model polyextremotolerant organism. These tools include those which can analyze base metabolism, such as Flux Balance Analysis (FBA) (Orth et al., 2010) and Flux Variability Analysis (FVA) (Steinn Gudmundsson & Thiele, 2010); tools which can aid in redesigning metabolism for optimization of a desired phenotype, such as OptKnock (Burgard et al., 2003) and OptForce (Burgard et al., 2003); and tools which elucidate potentially non-intuitive relationships in metabolism such as Flux Coupling Analysis (FCA) (Burgard, Nikolaev, Schilling, & Maranas, 2004). This work uses the standard measure of flux of mmol per gDW per h (Orth, Thiele and Palsson, 2010; Thiele and Palsson, 2010; Maranas and Zomorrodi, 2016). All optimization problems have primal and dual forms, both of which can be enlightening about the problem solution, particularly a quantity determined from the dual problem called the shadow price. The shadow price associated with variable i is defined as the reduction in the optimization objective caused by producing one more unit of i. Generally, shadow price is used in an economic sense to define the cost of some process in terms of currency; however, this metric can also be applied to the cost of some biological objective (e.g., growth) due to increasing production of a metabolite, such as a defensive pigment, by one unit. This can be determined using dual formulation of the

FBA problem. The cost of producing melanins and carotenoids by *E. dermatitidis* and the associated shadow prices, in particular, have not yet been investigated in this manner.

In this work, a draft GSM of *Exophiala dermatitidis* was first reconstructed from annotated genome of *E. dermatitidis* and an enzyme consensus between four GSMs from a related genus, Aspergillus, namely *A. niger* (Andersen et al., 2008), *A. nidulans* (David, Özçelik, Hofmann, & Nielsen, 2008), *A. oryzae* (Vongsangnak, Olsen, Hansen, Krogsgaard, & Nielsen, 2008), and *A. terreus* (J. Liu et al., 2013). Enzymes used in these Aspergillus GSMs (Andersen, Nielsen and Nielsen, 2008*;* David *et al.*, 2008; Vongsangnak *et al.*, 2008; Liu *et al.*, 2013) were used in conjunction with bidirectional BLASTp analyses to hypothesize characterizations of open reading frames. In general, the bidirectional BLASTp analyses assigned EC numbers, and the metabolic functionalities that accompany those numbers, to genes already annotated in the NCBI database with non-hypothetical protein names. This draft model next underwent manual and automated curation, the latter through using the tool OptFill (Schroeder & Saha, 2020a), to develop the *i*Ede2091 model. *i*Ede2091 was used in computational investigation of the metabolic cost of defensive pigment synthesis through shadow price analysis. This analysis shows that on both a per-carbon atom and a per-unit (monomer in the case of melanins and molecule in the case of carotenoids), carotenoids are more expensive to produce than melanins. Given that the functions of carotenoids and melanins are generally overlapping, this suggests that carotenoids perform a metabolically valuable protective role which has not been fully explored as of yet, potentially related to absorbance of violet and blue visible light. Finally, the potential of *Exophiala* as a model eumelanin-producing organism, particularly with respect to human eumelanin production in melanocytes, was investigated based on similarity of metabolic pathways and tyrosinase enzyme sequence similarity. This analysis showed that key amino acid residues are conserved in tyrosinase between *Exophiala* and humans, including residues whose mutations are associated with oculocutaneous albinism A1 (OCA1), which suggests *Exophiala* may be used as a model of human eumelanin-production.

3.3. RESULTS

3.3.1. Reconstruction of First Draft *E. dermatitidis* Model

In this work, the first draft GSM of *E. dermatitidis*, was reconstructed using logical Gene-Protein-Reaction (GPR) links to determine the set of metabolic reactions which occur in an organism using publicly available data such as NCBI and UniProt annotated genomes. This initial reconstruction was necessarily incomplete due to incomplete genome annotation, in that only approximately 43% of genes were annotated and less than 5% had some Enzyme Classification (EC) number annotation (*Exophiala dermatitidis NIH/UT8656 Genome Assembly*, 2011; *Exophiala dermatitidis (strain ATCC34100/CBS 525.76/NIH/UT8656)*, 2018). EC numbers were used to establish the GPR links, and therefore automated exploration of BRENDA was used to address this incompleteness and to retrieve more EC numbers, see Figure 3.1 and methods for more details of this procedure. From this, approximately 20% of genes were linked to some EC numbers. These proteins were then localized to their respective subcellular compartment through use of the CELLO subcellular localization tool (C. Yu & Lin, 2004), the results of which can be found in Supplemental Table S1 (see section 7.2 for how to access this file) This still left major metabolic gaps; therefore, in addition to genome annotation data, a core set of Enzyme Classification (EC) numbers were identified by being common to GSM models of four strains of a closely related genus (Aspergillus), *A. niger* (Andersen et al., 2008), *A. nidulans* (David et al., 2008), *A. oryzae* (Vongsangnak et al., 2008), and *A. terreus* (J. Liu et al., 2013), hereafter referred to as a the full consensus of *Aspergillus* enzymes. These *Aspergillus* models were chosen as they were the phylogenetically closest species

(Schoch et al., 2009) for which metabolic models were available. This work was limited to using the *Aspergillus* species models in that the next-closest fungi with GSMs published are at the phylum level, for example *Yarrowia* and *Saccharomyces* species which are quite phylogenetically distant. Further, all four *Aspergillus* species considered here have larger genome that *E. dermatitidis* allowing for greater genome coverage, while model Ascomycetes like *S. cerevisiae* and *Y. lipolytica* have smaller genomes. This restriction resulted in a more conservative metabolic reconstruction than might have otherwise been created in addition to limiting the number of reactions in the database for OptFill applications. This also limited the number of OptFill applications, as each new model considered would require one additional application. The full consensus of *Aspergillus* enzymes included 310 EC numbers in total. ECs already identified in *E. dermatitidis* were removed from the list of EC numbers belonging to the consensus of all four *Aspergillus* models (Andersen, Nielsen and Nielsen, 2008; David *et al.*, 2008; Vongsangnak *et al.*, 2008; Liu *et al.*, 2013), leaving 56 unique ECs. These 56 EC numbers were converted to metabolic functionalities and added to the existing draft model of *E. dermatitidis* as a set of functionalities likely common to these closely-related melanized fungal species (Schoch et al., 2009). See method and the GitHub "*E_dermatitidis*_model" repository (DOI: 10.5281/zenodo.3608172) for how this was accomplished. Steps taken in reconstruction can be found in greater detail in Supplemental Table S2.

3.3.1.1. Bidirectional BLASTp of Full Consensus Aspergillus Enzymes onto E. dermatitidis

The list of 56 ECs common to Aspergillus models but not identified in *Exophiala* were subjected to a bidirectional BLASTp against the *Exophiala* genome. This was accomplished through the Bidirectional BLAST Program (BBP) developed as part of this work, which can be found in the GitHub "E_dermatitidis_model" repository. The BBP program performs forward and backward BLASTp of amino acid sequences, taken from related species, encoding target ECs against a target genome in order to provide evidence for the presence of certain functionalities. The result of the BBP program when applied to the Aspergillus consensus ECs (Supplemental Table S3) was that 39 of 56 (69.6%) consensus ECs were identified in *Exophiala* with 169 unique bidirectional matches using conservative thresholds for the expect (1E-30) and percent positive substitution (60%) values. Many of these matches were between sequences annotated similarly in the reference Aspergillus species and *Exophiala*. Examples include annotations in *Aspergillus* species such as “xylulokinase”, “2-aminoadipate transaminase”, and “phosphoadenylyl-sulfate reductase (thioredoxin)” matching to annotations is *Exophiala* of “D-xylulose kinase A”, “aromatic amino acid aminotransferase I”, and “phosphoadenosine phosphosulfate reductase”, respectively. Other matches assigned EC number to multi-functional enzymes such as the “pentafunctional AROM polypeptide” being assigned to EC numbers 1.1.1.25, 2.5.1.19, and 4.2.1.10 based on strong sequence similarity to specific enzymes such as shikimate dehydrogenase, 3-phosphoshikimate 1-carboxyvinyltransferase, and 3-dehydroquinate dehydratase, respectively. In addition, a total of 22 bidirectional matches to protein sequences currently annotated as “hypothetical” proteins were made. These matches to hypothetical proteins mapped four hypothetical *Exophiala* protein sequences to seven EC numbers. The used reference *Aspergillus* sequences of six of these EC numbers, 1.2.1.38, 2.7.2.8, 6.3.3.1, 6.3.4.13, 6.3.4.14, and 6.4.1.2, only produced significant sequence alignment matched to hypothetical proteins in the *Exophiala* genome, indicating *in silico* identification of potentially unknown metabolic functionalities. Particularly important to this study is the identification EC 6.4.1.2, acetyl-CoA carboxylase, which produces malonyl-CoA. Malonyl-CoA is an essential precursor for the synthesis of hydroxylated naphthalene compounds which, when polymerized, produce DHN-melanin. See Figure 3.2 for DHN-melanin synthesis pathway with the reaction catalyzed by EC 6.4.1.2 which highlights the importance of this functionality.

3.3.1.2. From first draft E. dermatitidis model to second draft E. dermatitidis model

Despite the added functionality of the *Aspergillus* full consensus enzyme set and subsequent potential identification of new functionalities in the *Exophiala* genome, there were a number of “holes” in the metabolic reconstruction. These “holes” included lacking full synthesis pathways for defensive pigments and some biomass components. Therefore, the set of enzymes common to three of these Aspergillus models (Andersen, Nielsen and Nielsen, 2008; David *et al.*, 2008; Vongsangnak *et al.*, 2008; Liu *et al.*, 2013), the latest model of another ascomycete fungus, *Saccharomyces cerevisiae* (*i*Sce926) (Chowdhury et al., 2015), and literature information on fungal melanin synthesis (Paolo *et al.*, 2006; Chen *et al.*, 2014; Eisenman and Casadevall, 2012; Toledo *et al.*, 2017; Schmaler-Ripcke *et al.*, 2009), were used to manually address some metabolic gaps. Once this manual step was complete, the model could produce all required defensive pigments and biomass components and all Thermodynamically Infeasible Cycles (TICs) were addressed. In addition, the model was further refined to make sure that *Exophiala* can grow on carbon sources such as ethanol (J. Kumar, 2018), glucose (Poyntner *et al.*, 2016; Chen *et al.*, 2014), sucrose (Dadachova et al., 2007), and ethanol (J. Kumar, 2018) and to provide the opportunity to study metabolism, specifically pigment costs, under various different growth conditions. Once these objectives had been met, the resulting model was called the second draft *Exophiala* model. The second draft model had no TICs and consists of 1591 reactions, of which 711 could carry flux, and at best can produce 591 metabolites. For more details on the reconstruction of the first and second draft *Exophiala* models, see the methods.

3.3.1.3. From second draft E. dermatitidis model to iEde2091

The remainder of the set of enzymes common to three of four *Aspergillus* models was then converted to their metabolic functionalities (see methods), for a total of 344 reactions, and used as a database for the application of OptFill to the second draft model of *Exophiala*. Unfortunately, the large number of reactions in the model and database, as well as the large number of potential TICs between database and model, required several iterations of performing OptFill and removing from the database reactions participating in the most TICs identified in the allotted solution time (one week), until the database was reduced to 241 reactions, which allowed reasonable solution times (e.g. under 1 week to produce some gapfilling solutions). This procedure was repeated for the set of enzymes common to two of four Aspergillus models and to those unique to one model. This workflow is highlighted in Figure 3.1. In total, 43 reactions were added to the *Exophiala* model. This resulted in unblocking of a total of 82 reactions and 63 metabolites. Once each solution of this workflow was incorporated, the enzymes linked to filling solutions underwent a bidirectional BLASTp between reference *Aspergillus* sequences and the *Exophiala* genome, to determine the level of genomic support for these added reactions. This procedure was repeated for the set of enzymes common to two of four Aspergillus models and to the set of enzymes belonging to exactly one Aspergillus model. The resultant model was designated *i*Ede2091. The *i*Ede2091 model contains 1661 reactions (of which 824 can carry flux as determined by Flux Variability Analysis), 1856 metabolites, and 2091 genes. The set of genes includes those used to build the first draft model (861 genes) and those related to added metabolic functionality from the full consensus of Aspergillus model enzymes (33 genes), the set of enzymes common to three of four Aspergillus models (21 genes), the set of enzymes common to two of four Aspergillus models (2 genes), and the set of enzymes unique to an Aspergillus model (18 genes).

3.3.2. Applications of the *i*Ede2091 model

The *i*Ede2091 model was applied in two investigations. The first is the investigation of the shadow price of defensive pigments to better understand the costs and roles of the defensive pigments in polyextremotolerant systems. The second is the investigation of *Exophiala* melanin synthesis and comparison to that of humans to investigate the feasibility of using *Exophiala* as a model of human melanocytes.

3.3.2.1. Shadow price of defensive pigments and their precursors under various growth conditions

The *i*Ede2091 model was subjected to 36 growth conditions based on the available carbon source (sucrose, ethanol, acetate, or glucose), growth-limiting nutrient (carbon, nitrogen, or sulfur), and rate at which that limiting nutrient was made available to the system (low, moderate, or high). In this study, the growth-limiting nutrient or atom was defined as the nutrient which controls the rate of growth through its scarceness, while all other nutrients or atoms are provided in at least three order of magnitude excess. The rate of availability of the growth-limiting nutrient to the organism is also arbitrary because no information appears to be published which would suggest biologically relevant uptakes rates for *E. dermatitidis*. Shadow price is the change in the objective value of an optimization problem for one more unit of the desired product. As the model simulations were performed using the objective of maximizing biomass, all shadow prices are negative in value, and should be compared against a baseline growth rate of approximately 0.104 h^{-1} for non-stressed *Exophiala* growth in nutrient-limited conditions (Dadachova et al., 2007), since, as can be seen in Supplemental Table S4, the magnitude of the availability of the limiting resource and the growth rate, have no effect on the shadow price. Supplemental Table S4 shows that under arbitrarily defined high, medium, and low growth-limiting nutrient availability conditions (corresponding to high, medium, and low growth rates) the shadow price is constant. This was chosen as a baseline for comparison to shadow prices derived from the *i*Ede2091 model because no data is at present available to describe the rate of nutrient uptake by *E. dermatitidis* which would enable the use of *i*Ede2091 to estimate the growth rate. In the following analyses the per-atom rate of carbon uptake was standardized across the different carbon sources.

3.3.2.2. Carbon-limited conditions

Samples of shadow prices for melanins can be found in Figure 3.3(A) and 3.3(B). In general, DHN-melanin is more expensive than eumelanin and pyomelanin both on a per-carbon basis and a per-monomer basis. The higher per-monomer cost of DHN-melanin is due to both the higher per-carbon cost and monomers being composed of 10 carbons, as opposed to eight carbons for the other two types of melanin produced by *Exophiala*. As shown in Figures 3.3(A) and 3.3(B), not all carbon sources are equally effective in the production of melanins. Generally, melanins are most expensive, in terms of shadow cost, to produce when *Exophiala* is grown using sucrose as a sole carbon source, with the exception of producing eumelanin using acetate as a sole carbon source. For all cases, as suggested by the shadow prices in Figures 3.3(A) and 3.3(B), producing one additional $\text{mmol}\cdot\text{gDW}^{-1}\cdot\text{h}^{-1}$ of any melanin monomer would cause *Exophiala* to cease all growth, and even catabolize existing biomass to meet this demand.

In addition to investigating the pigments themselves, an investigation has been made into the shadow cost of precursor molecules to the pigments. Here, a precursor will be defined as molecules consumed by important enzymes related to pigment production or generally agreed upon as the metabolic branching point to pigment synthesis and all molecules “downstream” of that point. For instance, since tyrosinase is considered important in eumelanin synthesis, tyrosine and all molecules in eumelanin synthesis after tyrosine are considered eumelanin precursors. In this work, these pigment precursors have been included in Figures 3.2 and 3.4, and Supplemental Figures S1 and S2. With respect to the melanin precursors, per-carbon atom cost of the precursors is generally lower than that of the melanins that these produce. Further, precursor per-carbon atom shadow price is generally consistent from the point at which melanin synthesis pathways branch from other metabolic pathways (the branch point being the starting point of the syntheses depicted in Figures 3.2 and 3.4). One example can be clearly seen in the DHN-melanin synthesis pathway with 1,3,6,8-tetrahydroxynaphthalene, scytalone, and 1,3,8-trihydroxynaphthalene all having the same shadow cost. This consistency is not seen in those molecules more proximal to core metabolism such as acetate, ATP, CTP, and requisite amino acids to produce these precursors (such as tyrosine and cysteine). The shadow cost of melanin pigments and their precursors, are similar between ethanol and acetate growth conditions. This is because nearly the same set of reactions to metabolize both these carbon sources, with the primary difference being the generation of two molecules of NADH in the catalysis of ethanol to acetate. This has no effect on the shadow price of molecules such as tyrosinase, but has some effect in the shadow price of carotenoids (ethanol-grown *E. dermatitidis* has a lower shadow price for carotenoids, see Supplemental Figures S1 and S2). It can be noted that in the shadow prices of melanins and their precursors, these molecules are generally cheaper to produce when grown on sucrose or glucose substrates. This is primarily due to the fact that the precursors to tyrosine synthesis, namely d-erythrose-4-phosphate (with its own precursors of d-glyceraldehyde-3-phosphate and beta-d-fructose-6-phosphate) and phosphoenolpyruvate, are part of (or proximal to) the glycolysis/gluconeogenesis pathway. From sucrose or glucose, glycolysis is performed to produce these tyrosine precursors. On the other hand, from acetate and ethanol, gluconeogenesis is performed to produce these tyrosine precursors. Gluconeogenesis requires more energy than glycolysis to perform; therefore, the shadow cost of tyrosine-derived pigments is greater for *E. dermatitidis* when grown on acetate or ethanol in comparison to growth on sucrose or glucose.

The per-carbon atom shadow prices of the three carotenoids which are a part of *Exophiala* biomass as modeled in *i*Ede2091, namely β-carotene, β-apo-4'-carotenal, and neurosporaxanthin, are approximately equivalent, see Figure 3.3(C) and 3.3(D). Synthesis pathways used by *E. dermatitidis* to produce carotenoids, as well as the shadow prices of carotenoid precursors, can be found in Supplemental Figures S1 and S2. As the per-carbon atom shadow costs are approximately equivalent, the per-molecule differences in shadow cost are due to the difference in number of carbon atoms in the carotenoid molecules, as β-carotene contains 40 carbon atoms, whereas the other two carotenoid compounds contain 35 carbon atoms. Essentially, carotenoids are more expensive for the cell to produce than are melanins on a per-carbon atom basis.

3.3.2.3. Nitrogen- and Sulfur-limited conditions

The nitrogen source used by the model is ammonia and, as with the carbon-limited conditions, the availability of the growth-limiting nutrient has no effect on shadow cost. In this analysis, metabolites which do not contain nitrogen, including DHN melanin, pyomelanin, and all three investigated carotenoids, have no shadow cost under nitrogen-limited conditions. This makes sense in that all other atoms are provided to the system in excess; therefore, utilizing more of those excess atoms would not hamper biomass production. As such, the only melanin compound which has a shadow cost in these conditions is eumelanin, whose monomers contain a single nitrogen. In nitrogen-limited conditions, the per-nitrogen atom shadow cost is approximately 41 times higher than that of the per-carbon atom cost. The reasons for this are likely twofold. First, far less nitrogen is needed by *Exophiala* to produce biomass than carbon (approximately 9.1:1 C:N in the biomass pseudomolecule). Second, not all nitrogen uptaken can be used by *Exophiala*, and utilization of nitrogen is less efficient than utilization of carbon. For instance, waste nitrogen is excreted in a nitrogen compound containing four nitrogen atoms (in urate), as opposed to the majority of waste carbon being expelled as carbon dioxide.

Similarly, in the cases where sulfur is the nutrient limiting model growth, compounds which contain no sulfur atoms have no shadow cost, including all defensive pigments studied. Therefore, only melanin precursors have a shadow cost under these conditions, which includes coenzyme A (CoA), its precursors, and all molecules containing CoA such as malonyl-CoA and acetyl-CoA. These compounds have relatively high per-sulfur atom shadow costs, of -28.73 h^{-1}, since each mole of the biomass pseudomolecule contains approximately 0.035 sulfur atoms, indicating that the sulfur needs of *Exophiala* are very low. Therefore, to produce one extra mmol·gDW^{-1}·h^{-1} of a sulfur-containing compound, a large amount of biomass would need to be catabolized.

3.3.2.4. Comparison of human and E. dermatitidis melanin synthesis

The melanin synthesis pathway of *Exophiala* and humans was compared in two ways: first by the series of reactions which produce human melanins (namely pheomelanin and eumelanin, see Figure 3.4), and second by comparison of the tyrosinase enzymes (see Figure 3.5).

In building the *i*Ede2091 model, we recognized that fungal melanins are typically transported in exocytic vesicles to the cell surface, where they are then attached to the cell wall (Camacho *et al.*, 2019; Upadhyay *et al.*, 2016). This pathway shares features with that observed in melanocytes, whereby synthesis occurs in specialized melanosomes. Moreover, the *Exophiala* and human pathways to produce the indole-5,6-quinone monomer of eumelanin are identical. Further, the production of pheomelanin in humans appears replicable in *Exophiala* should cysteine be added to the extracellular environment. The 5,6-indolequinone-2-carboxilic acid eumelanin monomer is not producible by *Exophiala* due to its lack of a tyrosinase-related protein. In investigating the potential for *Exophiala* to produce pheomelanin, the shadow price of cysteine was also investigated. With respect to carbon-limited conditions (see Figure 3.4), cysteine is more expensive than most other precursors on a per-carbon basis, particularly in cases of growth on sucrose and glucose. With respect to nitrogen-limited growth cases (see Supplemental Table S4), cysteine is very similar in cost to other amino acids. With respect to sulfur-limited growth cases (see Supplemental Table S4), the per-sulfur atom cost is high (around 28 h^{-1}), and is similar in cost to coenzyme A.

The four tyrosinase gene copies in *E. dermatitidis* where identified through genome annotation. Further, these sequences where used as a BLASTp query against the *E. dermatitidis* genome to confirm that these four were the only tyrosinase gene copies in *E. dermatitidis*. A non-redundant BLASTp analysis was performed by using the tyrosinase amino acid sequences of *Exophiala* as the search sequence against the human genome to determine the sequence similarity. This produced no matches of acceptable expect value (e.g. less than 1E-10), indicating large sequence dissimilarities. However, a COBALT alignment (Papadopoulos & Agarwala, 2007) of the amino acid sequences of three human tyrosinase alleles, human tyrosinase-related proteins, and the four gene copies of *Exophiala* produces more nuanced results. Tyrosinase-related proteins (TYRPs) have the same evolutionary origin as tyrosinase and are still very similar and were therefore included in this analysis (Furumura et al., 1998). The major catalytic difference between TYRPs and tyrosinases is that they act upon L-Dopachrome differently, one producing 5,6-indolequinone-2-carboxylic acid eumelanin monomers and the other producing indole-5,6-quinone eumelanin monomers. Portions of this alignment, namely the sequences related to the Copper binding domains A (CuA) and B (CuB) which constitute the active side of tyrosinase, are shown in Figure 3.5 using the 3-bit highlighting method. This method highlights in red aligned residues which have the same or very similar chemical structure, in blue somewhat conserved regions, and in grey unconserved regions. When highlighting key structural (orange triangle), functional (brown triangle), and active site (purple triangle) residues, it appears that these key residues are highly conserved between human tyrosinase-related proteins and tyrosinase and *Exophiala* tyrosinases. Poor BLASTp alignment scores appear to be due to substitutions, deletions, or lack of sequence conservation of non-critical residues, gaps in less critical regions of tyrosinase (such as residues which are not a part of secondary structures, such as the gap in CuA), and significant differences in enzyme length. This is shown in the Multiple Sequence Alignment (MSA) view shown in Figure 3.5. As an example of the length differences, while human tyrosinase has a primary structure of 529 amino acids, and tyrosinase-related proteins 1 and 2 have structures of 537 and 519 amino acids, respectively, while *Exophiala* tyrosinase lengths range from 381 to 614 amino acids. Much of the differences in length are in those sequences upstream of CuA and downstream of CuB, see the sequence identity summary shown in Figure 3.5. Interestingly, Figure 3.5 highlights residue mutations which trigger the switch between tyrosinase and tyrosinase-related proteins (pink triangles at 214, 219, 389, and 393 (García-Borrón & Solano, 2002)). In some gene copies of *Exophiala* tyrosinase, particularly the copy labeled as "Ede_un1", key residues which when mutated cause the switch between tyrosinase and tyrosinase-related protein are particularly well conserved, suggesting that *Exophiala* could be engineered to have a tyrosinase-related protein. Should this additional monomer synthesis pathway be engineered in *Exophiala*, through gene insertion or selective mutation, the melanin synthesis pathways between *Exophiala* and humans could be very similar. In addition to this analysis, a sequence alignment analysis to the Hidden Markov Model (HMM) using the Pfam tool (El-Gebali et al., 2019) was performed. This tool acknowledged the strong sequence similarity of *E. dermatitidis* tyrosinase enzymes with that of the general pattern of tyrosinase enzymes. The results of this analysis can be found in Supplemental Data S1.

In considering the uses of *Exophiala* as a model system of human melanin production, some amino acid residue positions where residue substitutions are associated with oculocutaneous albinism A1 (OCA1), which accounts for approximately 50% of cases of albinism worldwide and is caused by a non-functional tyrosinase in humans (Kamaraj & Purohit, 2014), are shown in black rectangles (Spritz, 1994) to highlight the potential for *Exophiala* as a model system to study OCA1.

3.4. DISCUSSION

In this work, a stoichiometric GSM of *E. dermatitidis* (*i*Ede2091) consisting of 1661 reactions, 1856 metabolites, and 2091 genes was developed in order to investigate *Exophiala* as a potential model organism for extremotolerant fungi and human melanocytes. Several issues were encountered in the metabolic reconstruction. First, the low levels of genome annotation (43%

annotated but less than 5% with associated enzyme classifications) represented knowledge gaps in the understanding of *Exophiala* metabolism that lead to many gaps and blocked reactions throughout the stages of reconstruction. This was dealt with by using four metabolic models from the related Aspergillus genus (Andersen, Nielsen and Nielsen, 2008; David *et al.*, 2008; Vongsangnak *et al.*, 2008; Liu *et al.*, 2013) in addition to the OptFill tool (Schroeder & Saha, 2020a) for TIC-free gapfilling of models, see Figure 3.1. The low levels of genome annotation also hindered the ability to create gene-protein-reaction links, which was addressed by using *Aspergillus* protein sequences as enzyme reference sequences for use in BLASTp analyses. This resulted in a large number of previously annotated genes being linked with enzyme classifications, and the functional identification of four sequences which may not yet have been identified.

In the shadow price investigation of melanins (Figure 3.3), it was noted that DHN-melanin has a higher per-unit cost than other melanins. This appears to be due simply to the larger number of carbon molecules present in each monomer unit when comparted to other melanins (see Figures 3.2 and 3.4). Furthermore, the difference between DHN-melanin, eumelanin, and pyomelanin in media where sucrose is the limited carbon source is that the latter two are synthesized from l-tyrosine, whereas DHN-melanin is synthesized from malonyl-CoA. The higher shadow price appears to be due to the higher per-carbon atom cost to produce acetate from sucrose which is perpetuated through the DHN-synthesis pathway. As shown in Figure 4.3, both melanin and carotenoid pigments are "cheapest" to produce in carbon-limited cases when glucose is the carbon source. This is due to the lack of preprocessing needed (e.g. other carbon sources may require gluconeogenesis or other metabolic transformations before being shunted to major energy-harvesting pathways).

The changing shadow prices for these defensive pigments under different growth conditions suggest that the profile of pigments (i.e., the type and quantity of defensive pigments) as produced by *Exophiala* varies by nutrient availability. In other words, the "cheaper" defensive pigments may be produced more than the expensive pigments. Having a range of defensive pigments (e.g. three melanin types and various carotenoids) with differing synthesis pathways makes them to be more or less expensive depending on available nutrients. This, in turn, may help minimize the cost of the extremotolerant nature of *Exophiala* by allowing the organism to preferentially produce the least expensive defensive pigment(s). The relatively high fractions of biomass accounted for by defensive pigments, 1.3 wt% for melanin (Philip Anthony Geis, 1981) and 3.5 wt% for carotenoids (Strobel, Breitenbach, Scheckhuber, Osiewacz, & Sandmann, 2009), as well as their high shadow prices suggest that these pigments are continually produced and stockpiled because increasing production of these pigments to meet cell need if the environment were to quickly become extreme is untenable.

The higher per-carbon shadow prices of carotenoids compared to melanins might help to expand the current understanding of the role of carotenoids. Firstly, carotenoids are a secondary line of defense against external extreme conditions as they are deposited in the cell membrane (Chen *et al.*, 2014; Kumar *et al.*, 2018), whereas melanins are deposited in the cell wall (Chen *et al.*, 2014; Geis, 1981; Szaniszlo, 2002). Secondly, melanins are known to provide protection against antifungal and antimicrobial compounds (Paolo *et al.*, 2006; Toledo *et al.*, 2017; Nosanchuk and Casadevall, 2006b); lytic enzymes (Paolo et al., 2006); heat and cold stress (Paolo *et al.*, 2006; Toledo *et al.*, 2017); rapid freezing (Paolo et al., 2006); ionizing radiation (Kumar, 2018; Dadachova *et al.*, 2007); oxidative stress (Toledo et al., 2017); UV radiation (Toledo et al., 2017); heavy metals (Kumar, 2018; Singh *et al.*, 2013); light (Z. Chen et al., 2014); and immune responses (Z. Chen et al., 2014). Further, several genes related to both eumelanin and DHN-melanin synthesis are upregulated under low pH stress, suggesting that melanins are also produced under pH stress (Z. Chen et al., 2014). At present, it is known that carotenoids protect against stress conditions such as oxidative stress (such as free radicals) (Kumar *et al.*, 2018; Strobel *et al.*, 2009; Avalos and

Carmen Limón, 2015), UV radiation (Geis and Szaniszlo, 1984; Kumar, 2018; Strobel *et al.*, 2009; Avalos and Carmen Limón, 2015), and light (Kumar, 2018; Strobel *et al.*, 2009; Avalos and Carmen Limón, 2015). Each function of carotenoids is already accounted for by melanins. It has been suggested that carotenoids do not play a physiological role in fungi, but rather function as precursors to the synthesis of other biomolecules (Avalos & Carmen Limón, 2015). However, this appears inconsistent with their higher shadow cost in comparison to melanin compounds which can accomplish the same functions with deposition in the cell membrane and high weight fraction in some fungal species (Strobel et al., 2009). Although several previous works hinted about the possibility of carotenoids having unexplored functions in fungi (Chen *et al.*, 2014; Avalos and Carmen Limón, 2015), this is the first study that provides a computational and systems biology perspective. One study has postulated that perhaps carotenoids protect against light which passes through the melanin in the cell wall (Z. Chen et al., 2014). This seems a likely function as melanin absorbance of electromagnetic radiation is high in the ultraviolet spectrum to approximately 400 nm in wavelength, and exponentially declines in the wavelength range of 400 to 500 nm (Of *et al.*, 2015; Ou-Yang, Stamatas and Kollias, 2004), whereas this latter range constitutes the peak absorbance of carotenoids (Yamamoto and Bangham, 1978; Zaghdoudi *et al.*, 2017). Thus, the combination of these two pigments would protect *Exophiala* cell from the UV spectrum through higher-energy visible light (namely violet and blue light). The high cost of producing carotenoids along with high fraction of cell weight does suggest that the violet and/or blue light is particularly disruptive to some high-value metabolic process in *Exophiala* which should be further investigated.

In exploring the suitability of *Exophiala* as a model organism for human melanocytes, the sequence alignment results of *Exophiala* and human tyrosinase enzymes show that CuB is the best-conserved portion of tyrosinase active site, through all key amino acids, and therefore likely the essential structures of CuA is also preserved. As tyrosinase is the key enzyme in eumelanin synthesis in both *E. dermatitidis* and human, several residues associated with OCA1 are persevered between the species. Since *Exophiala* has a significantly smaller genome (26.4 Mb compared to 3253.8 Mb for human), *Exophiala* may be used as a model system for human eumelanin production. As OCA1 is the most prevalent type of albinism worldwide, *Exophiala* may be used as a model system for studying causal mechanisms of OCA1 and potentially to identify treatment options. Unfortunately, African populations, where albinism is a more pressing social and health problem (Brilliant, 2015), would benefit less from *Exophiala* eumelanin studies, than for Caucasian and Asian populations, as approximately 77% of albinism cases in African populations result from oculocutaneous albinism A2 (OCA2), with most of the remainder is attributed to OCA1. OCA2 is a result of the lack of a tyrosinase transporter proteins, called P protein which is necessary to transport tyrosinase into human melanosomes (a subcellular compartment dedicated to melanin synthesis in melanocytes) and/or stabilize tyrosinase (Kamaraj & Purohit, 2014), which was not identifiable through *in silico* methods in *Exophiala*, such as through BLASTp or annotated genomes. Therefore, further study of *Exophiala* is warranted to identify this transporter protein and improve the potential for *Exophiala* as a model system.

Furthermore, it was determined that one type of melanin produced by human, is not produced by *Exophiala*. Pheomelanin is a red-brown to yellow pigment (S. Ito & Wakamatsu, 2011), and it is likely that *Exophiala* could produce this type of melanin. No additional enzymes are needed to produce pheomelanin beyond that which *Exophiala* already possesses (see Figure 4.4), rather free cysteine in the location of eumelanin synthesis is required (Shosuke Ito, 2003). Growing *Exophiala* in a cysteine-rich culture or engineering a cysteine pump to the extracellular space of *Exophiala* could result in pheomelanin production, allowing production of both human melanin types. Alternatively, if *Exophiala* were to provide the cysteine for pheomelanin synthesis, given its high shadow price and the shadow price of dopaquinone, it is reasonable to hypothesize that the resultant pheomelanin would be the costliest melanin produced by *E. dermatitidis*. This is perhaps why *E. dermatitidis* does not natively produce pheomelanin.

Overall, the results of this work suggest several potential interesting *in vivo* follow-up studies that will increase our understanding of extremotolerant fungi using *Exophiala* as a model system. Key predictions arising from the *i*Ede2091 model that are currently being tested include assessing the effects of different carbon sources on melanin and carotenoid accumulation, as well as determining the effects of mutations that abrogate specific metabolic pathways on pigment production. In addition, phenotypic profiling of mutants defective in melanin and/or carotenoid synthesis is underway to better evaluate the roles of each pigment in stress tolerance. Although detailed *in vivo* investigation may be needed to further establish *Exophiala* as a potential model organism for human melanocytes including demonstrating the production of pheomelanin, this work attempts to enable a broader understanding of melanin production across kingdoms.

3.5. FIGURES

See next page.

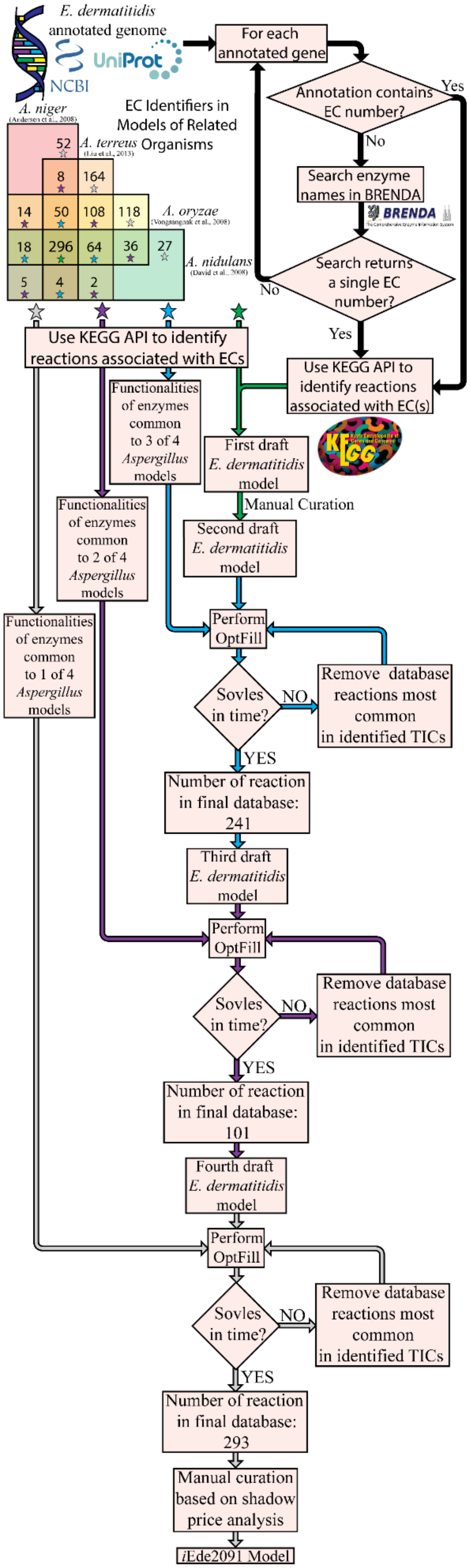


Figure 3.1: Workflow of iEde2091 GSM reconstruction.

Extended Caption: This figure shows the reconstruction workflow of *i*Ede2091, beginning with the annotated genomes from NCBI and UniProt. These gene names taken from these annotated genomes are then used to automatically search the BRENDA database for the associated Enzyme Classification (EC) number. This data was combined with the consensus of enzymes present in the selected *Aspergillus* species GSM reconstructions to form the first draft *E. dermatitidis* GSM model. After manual curation to ensure production of defensive pigments and biomass, this became the second draft *E. dermatitidis* model. Subsequent draft *E. dermatitidis* models were created by using the OptFill tool to fill metabolic gaps using non-consensus *Aspergillus* databases. Once each non-consensus database had been used, the *i*Ede2091 model was complete.

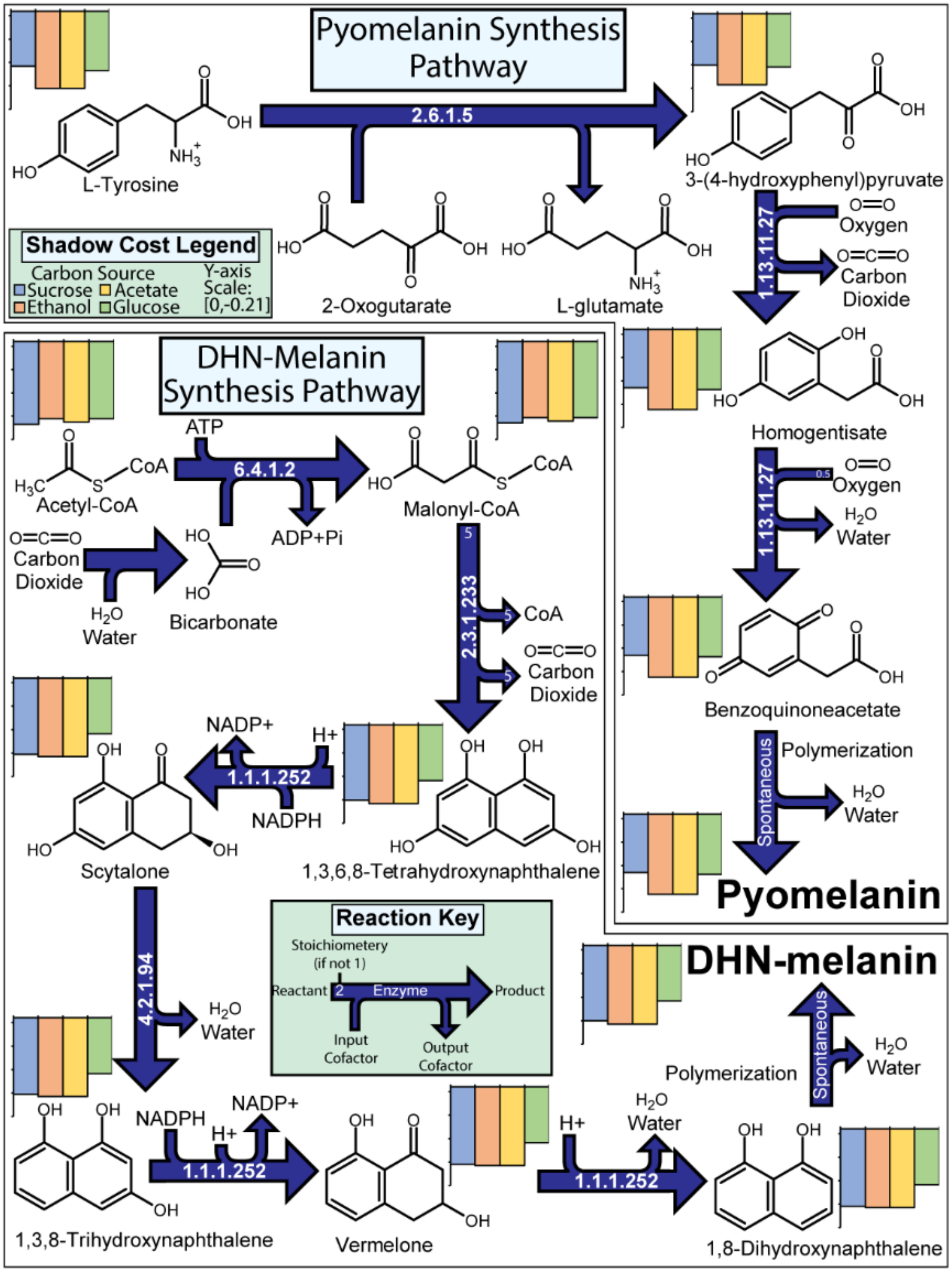


Figure 3.2: Synthesis pathways of pyomelanin and DHN-melanin in E. dermatitidis.

Extended Caption: This figure shows the synthesis pathways of pyomelanin and DHN-melanin including chemical structures, reaction stoichiometeries, catalyzing Enzyme Classification (EC) number, and reaction cofactors.

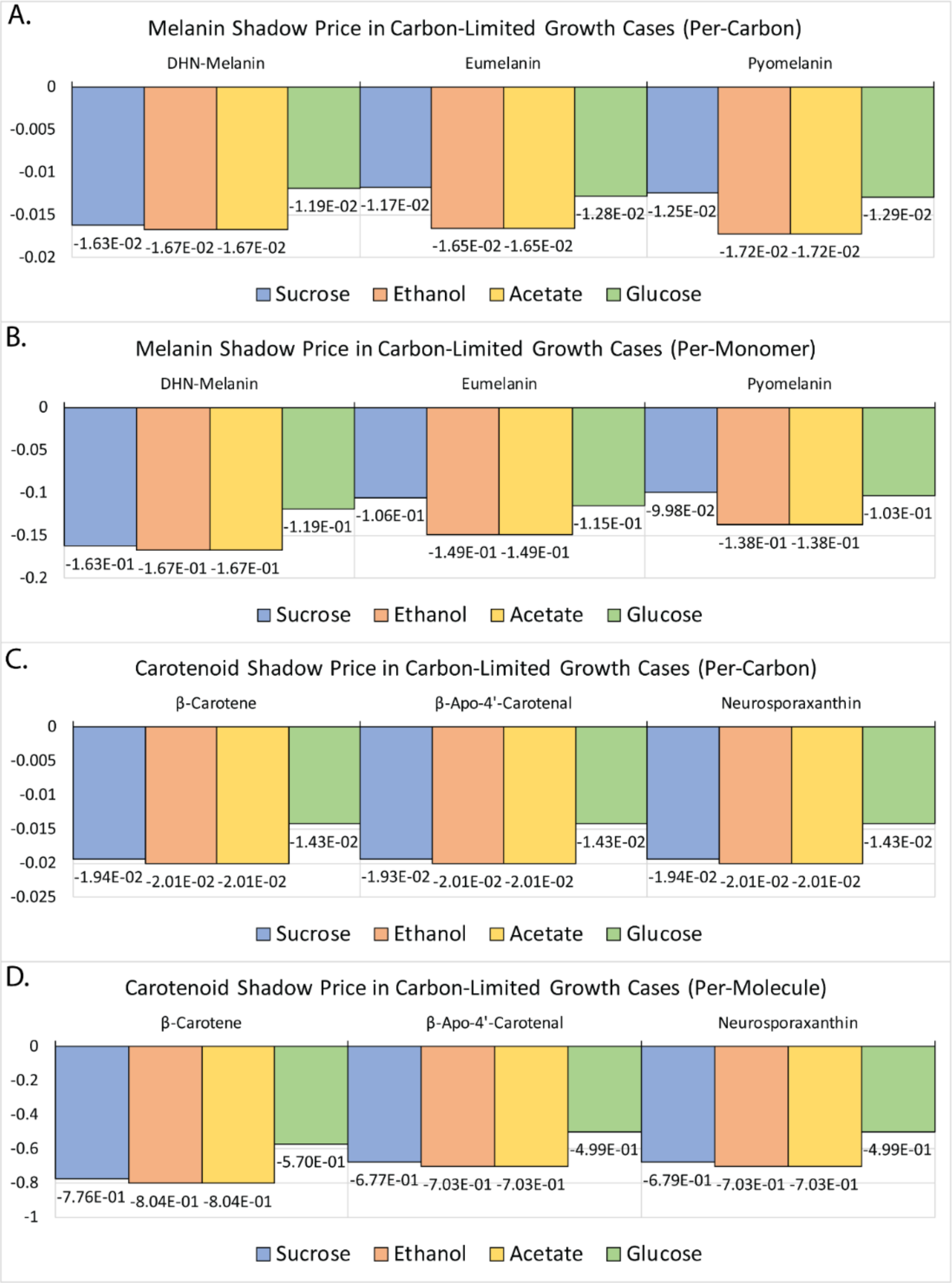


Figure 3.3: Shadow prices of E. dermatitidis pigments.

Extended Caption: This figure shows bar graphs of *E. dermatitidis* defensive pigment shadow prices under carbon-atom limited conditions, using four different carbon sources, on per-limited atom and per-unit basis. (A) Per-carbon atom shadow costs of the three melanins producible by E. dermatitidis under various carbon-limited growth conditions. (B) Per-monomer shadow costs of the three melanins producible by E. dermatitidis under various carbon-limited growth conditions.

(C) Per-carbon atom shadow costs of the three carotenoids which are modeled to constitute E. dermatitidis biomass under various carbon-limited growth conditions. (D) Per-molecule shadow costs of the three carotenoids which are modeled to constitute E. dermatitidis biomass under various carbon-limited growth conditions.

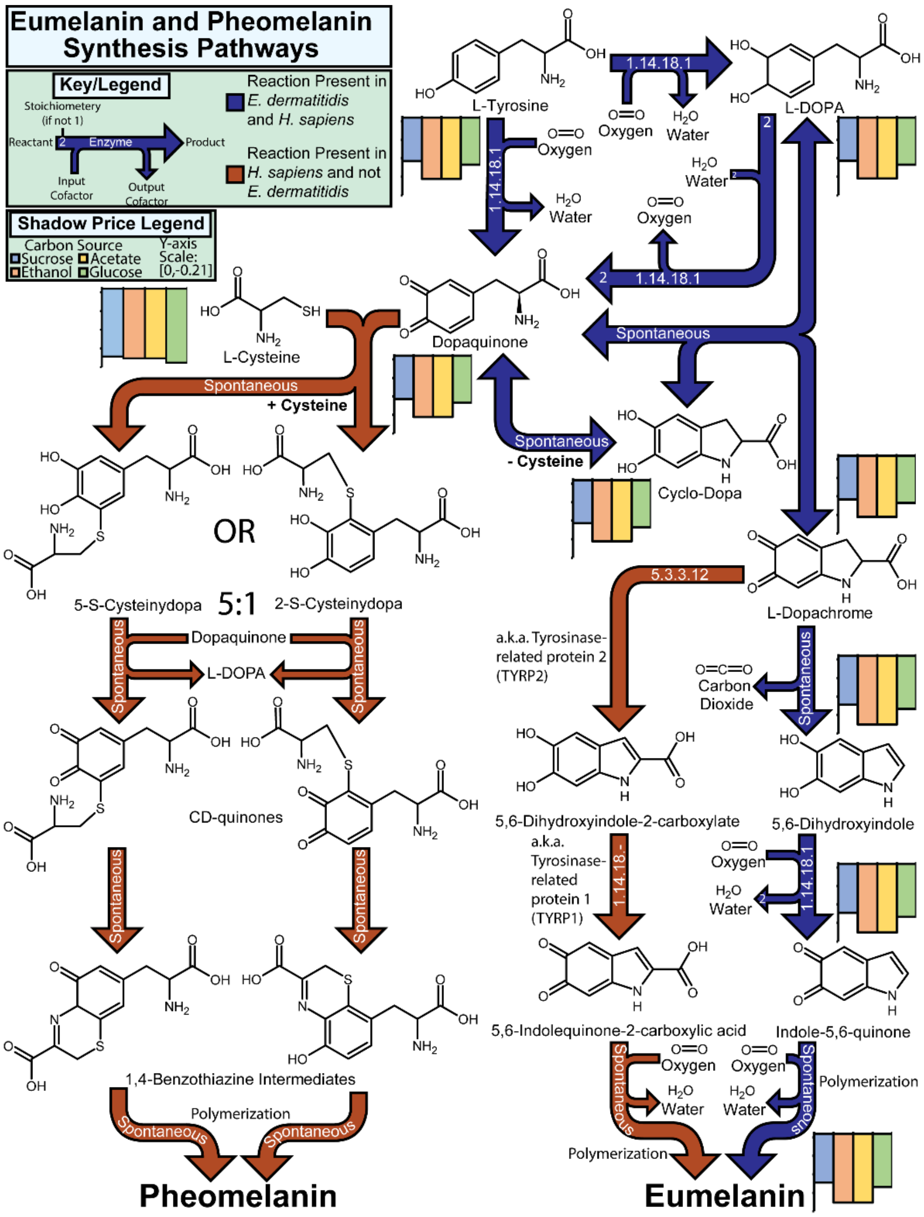
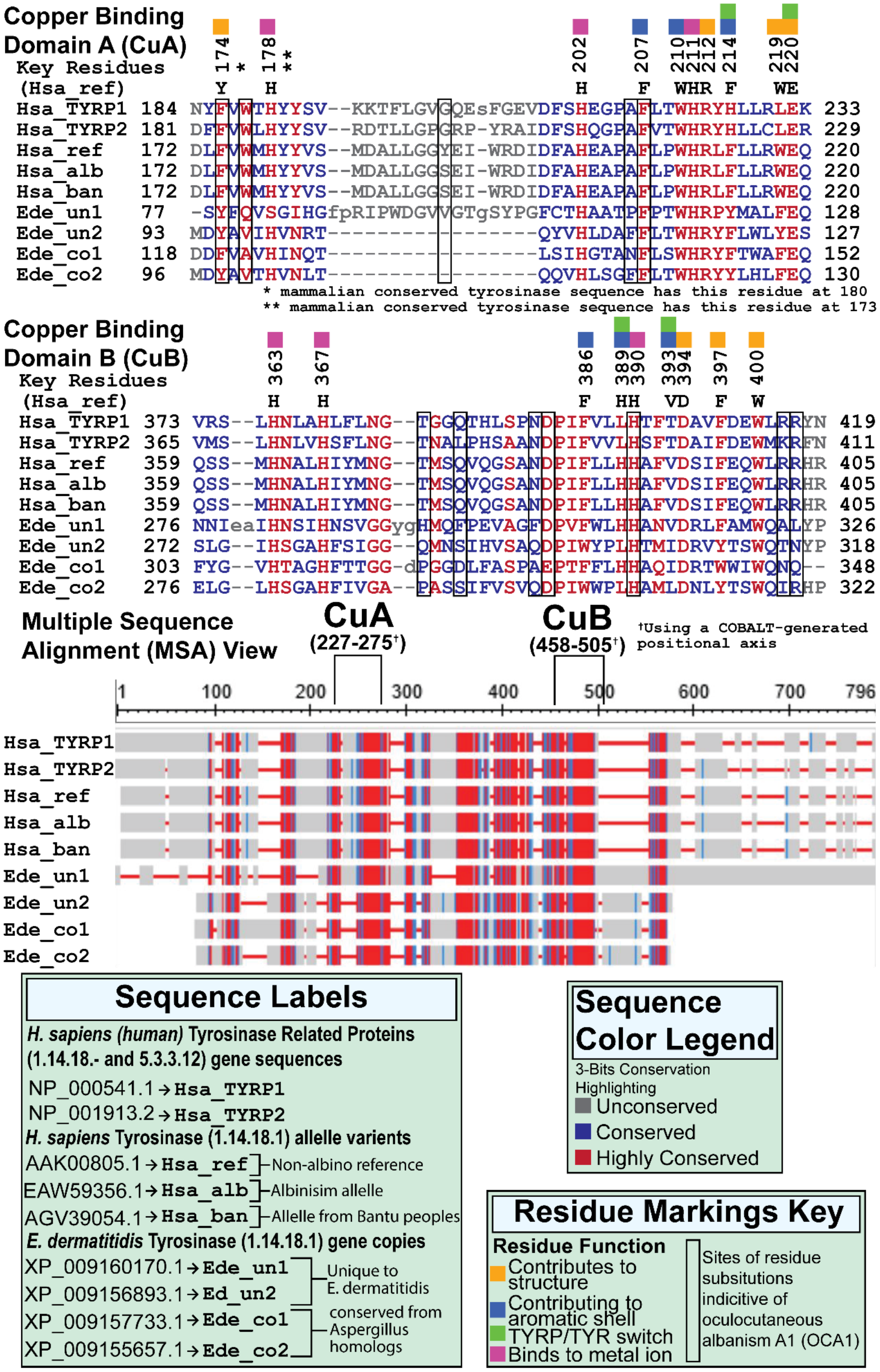


Figure 3.4: Synthesis pathways of eumelanin and pheomelnin in humans and E. dermatitidis.

Extended Caption: This figure shows the synthesis pathways of eumelanin and pheomelanin including chemical structures, reaction stoichiometeries, catalyzing Enzyme Classification (EC) number, and reaction cofactors in humans (green and blue arrows) and E. dermatitidis (blue arrows). The major difference between these species' eumelanin synthesis pathways is the presence of tyrosine-related proteins (TYRPs) in humans which catalyze the reactions indicated by green arrows. In both species, the key initiating enzyme is tyrosinase, Enzyme Classification 1.14.18.1 which catalyzes the initial steps of eumelanin synthesis. A deficiency in tyrosinase activity may result in oculocutaneous albinism A1 in humans. The second type of human melanin, pheomelanin, is largely produced by spontaneous reactions beyond the tyrosinase-catalyzed production of dopaquinone. The branching of eumelanin and pheomelanin production is accomplished by the presence or absence of cysteine where dopaquinone is concentrated. This suggests that pheomelanin may be inducible in *E. dermatitidis*.



```
Hsa_TYRP1 184 NYFVWTHYYSV--KKTFLGVGQEsFGEVDFSHEGPAFLTWHRYHLLRLEK 233
Hsa_TYRP2 181 DFFVWLHYYSV--RDTLLGPGRP-YRAIDFSHQGPAFVTWHRYHLLCLER 229
Hsa_ref   172 DLFVWMHYYVS--MDALLGGYEI-WRDIDFAHEAPAFLPWHRLFLLRWEQ 220
Hsa_alb   172 DLFVWMHYYVS--MDALLGGSEI-WRDIDFAHEAPAFLPWHRLFLLRWEQ 220
Hsa_ban   172 DLFVWMHYYVS--MDALLGGSEI-WRDIDFAHEAPAFLPWHRLFLLRWEQ 220
Ede_un1   77  -SYFQVSGIHGfpRIPWDGVVGTgSYPGFCTHAATPFPTWHRPYMALFEQ 128
Ede_un2   93  MDYAVIHVNRT------------------QYVHLDAFFLTWHRYFLWLYES 127
Ede_co1   118 DDFVAVHINQT------------------LSIHGTANFLSWHRYFTWAFEQ 152
Ede_co2   96  MDYAVTHVNLT------------------QQVHLSGFFLTWHRYYLHLFEQ 130
```

* mammalian conserved tyrosinase sequence has this residue at 180
** mammalian conserved tyrosinase sequence has this residue at 173

```
Hsa_TYRP1 373 VRS--LHNLAHLFLNG--TGGQTHLSPNDPIFVLLHTFTDAVFDEWLRRYN 419
Hsa_TYRP2 365 VMS--LHNLVHSFLNG--TNALPHSAANDPIFVVLHSFTDAIFDEWMKRFN 411
Hsa_ref   359 QSS--MHNALHIYMNG--TMSQVQGSANDPIFLLHHAFVDSIFEQWLRRHR 405
Hsa_alb   359 QSS--MHNALHIYMNG--TMSQVQGSANDPIFLLHHAFVDSIFEQWLRRHR 405
Hsa_ban   359 QSS--MHNALHIYMNG--TMSQVQGSANDPIFLLHHAFVDSIFEQWLRRHR 405
Ede_un1   276 NNIeaIHNSIHNSVGGygHMQFPEVAGFDPVFWLHHANVDRLFAMWQALYP 326
Ede_un2   272 SLG--IHSGAHFSIGG--QMNSIHVSAQDPIWYPLHTMIDRVYTSWQTNYP 318
Ede_co1   303 FYG--VHTAGHFTTGG-dPGGDLFASPAEPTFFLHHAQIDRTWWIWQNQ-- 348
Ede_co2   276 ELG--LHSGAHFIVGA--PASSIFVSVQDPIWWPLHAMLDNLYTSWQIRHP 322
```

Sequence Labels

H. sapiens (human) Tyrosinase Related Proteins (1.14.18.- and 5.3.3.12) gene sequences
NP_000541.1 → Hsa_TYRP1
NP_001913.2 → Hsa_TYRP2
H. sapiens Tyrosinase (1.14.18.1) allele varients
AAK00805.1 → Hsa_ref – Non-albino reference
EAW59356.1 → Hsa_alb – Albinisim allele
AGV39054.1 → Hsa_ban – Allele from Bantu peoples
E. dermatitidis Tyrosinase (1.14.18.1) gene copies
XP_009160170.1 → Ede_un1
XP_009156893.1 → Ed_un2 – Unique to E. dermatitidis
XP_009157733.1 → Ede_co1
XP_009155657.1 → Ede_co2 – conserved from Aspergillus homologs

Sequence Color Legend

3-Bits Conservation Highlighting
■ Unconserved
■ Conserved
■ Highly Conserved

Residue Markings Key

Residue Function
■ Contributes to structure
■ Contributing to aromatic shell
■ TYRP/TYR switch
■ Binds to metal ion
Sites of residue subsitutions indicitive of oculocutaneous albanism A1 (OCA1)

Figure 3.5: Tyrosinase and tyrosinase-related protein sequence alignments between humans and E. dermatitidis.

Extended Caption: This figure shows portions of the sequence alignments performed by NCBI's COBALT tool using the amino acid sequences of human tyrosinase-related protein 1 (Has_TYRP1, accession NP_000541.1), 2 (Has_TYRP2, accession NP_01913.2), a reference allele human tyrosinase sequence (Has_ref, accession AAK00805.1), an oculocutaneous albinism A1 allele (Has_alb, accession EAW59356.1), an allele from an individual of the Bantu peoples (Has_ban, accession AGV39054.1), and reference sequences for the four tyrosinase gene copies of *E. dermatitidis* (Ede_un1, accession XP_009160170.1; Ede_un2, accession XP_009156893.1; Ede_co1, accession XP_009157733.1; and Ede_co2, accession XP_009155657.1). The portions of the alignments shown concern the two parts of the active site of tyrosinase, Copper Binding Domains A and B (CuA and CuB, respectively). It is shown that all amino acid residues thought to be critical to active site function (key residues) are highly conserved between the two7 species (García-Borrón & Solano, 2002). Further, many sites where amino acid substitutions are associated with oculocutaneous albinism A1 (residues boxed in black) are conserved between species. Also shown is the Multiple Sequence Alignment (MSA) view, which shows that the active sites and many sequences between the active sites are preserved between the aligned sequences. This further shows that the large differences in sequence lengths between genes are largely due to sequences flanking the active sites.

3.6. MATERIALS AND METHODS

3.6.1. Use of E. dermatitidis genome annotation information

Genome annotations of *E. dermatitidis* were retrieved from NCBI (“Exophiala dermatitidis NIH/UT8656 Genome Assembly,” 2011) and UniProt (“Exophiala dermatitidis (strain ATCC34100/CBS 525.76/NIH/UT8656),” 2018) databases. These initial retrievals contained 9562 and 9391 Open Reading Frames (ORFs), respectively, with up to 4.9% of ORFs labeled with Enzyme Classification (EC) numbers. As EC numbers are often used to establish Gene-Protein-Reaction (GPR) links in a GSM, programming scripts and a library of programming functions were devised to automatically search the BRENDA database with the protein name to attempt to discover EC numbers for as many proteins as possible. These can be found in the GitHub “E_dermatitidis_model” repository which accompanies this work. This resulted in 2020 (21.5%) and 1724 (18.0%) of UniProt and NCBI ORFs, respectively, that correspond to at least one specific EC number. Accepting only single EC number BRENDA search results (to discount matches due to ambiguous names), and only those which are present in both annotations resulted in 532 EC numbers. Using these EC numbers, the reactions which KEGG indicated could be catalyzed by these EC numbers (determined by File SFF) were used to form the first draft model. These reactions were assigned to subcellular compartments by inputting the FASTA corresponding to the 537 EC numbers to the CELLO predictor for subcellular localization (C. Yu & Lin, 2004)(C. S. Yu, Lin, & Hwang, 2006). These results, included in Supplemental File 4, predicted 533 cytosolic, 435 mitochondrial, 66 extracellular, 6 lysosomic, 117 peroxisomic, 144 nucleic, and 5 endoplasmic reticulate reactions. Cytosolic, mitochondrial, and extracellular reactions were selected for incorporation to the model. Peroxisomic reactions were not included due to large number of metabolic gaps, resulting in many reactions which produced and/or consumed metabolites not present elsewhere in *Exophiala* metabolism, and for which literature evidence justifying their inclusion could not be found. Further, the lack of information in literature as to the metabolism which occurs in fungal peroxisomes and metabolite transporters further hinders accurate reconstruction of peroxisome metabolism to the extent that accurate reconstruction may not be possible at present. Nucleic reactions were not included as most reactions in this organelle involve the synthesis, breakdown, modification, or maintenance of RNA and DNA, and do not generally participate in other metabolic processes. Further, as with the peroxisome, some metabolites were present here which were not present elsewhere in *Exophiala* metabolism (other than DNA/RNA). The mitochondrial compartment was separated to inner and outer mitochondria, resulting in four distinct model compartments. The outer mitochondrion is modeled as compartment to store protons pumped by oxidative phosphorylation and another biologically relevant membrane across which transport must occur. The set of cytosolic, mitochondrial, and extracellular reactions were used in the definition of the first draft model of *Exophiala*.

3.6.2. Choice of Aspergillus models for metabolic gapfilling

Utilizing the most comprehensive phylogenetic tree for the Ascomycota phylum found (Schoch et al., 2009), the phylogenetic branches of the Ascomycota phylum where investigated for Genome-Scale Models (GSMs) which were created for related organisms (nearest branches were investigated first). The most closely related models identified belong to the *Aspergillus* genus, namely: *A. niger* (Andersen et al., 2008)*, A. nidulans* (David et al., 2008)*, A. oryzae* (Vongsangnak et al., 2008)*,* and *A. terreus* (J. Liu et al., 2013) which all belong to the same class as *E. dermatitidis*, Eurotiomycetes. No other genome-scale models belonging to this same class were identified. Genome-Scale models belonging to the same phylum as *E. dermatitidis* were considered for inclusion in the definition of automated gapfilling database, but this was dismissed for multiple reasons. First, this would result in a linear increase with the number of included models in the number of OptFill runs needed based on commonality of enzymes, resulting in a less tractable study. Second, this would reduce the number of core enzymes to those common to the phylum as opposed to the class. Third, this allows for a more conservative metabolic reconstruction, reducing the chances of adding false functionalities. Fourth, manually completed BLASTp analyses between *E. dermatitidis* and the model Ascomycete, *Saccharomyces cerevisiae*, showed poor alignments between sequences encoding the same enzyme. Fifth, there are acknowledged conserved homologs between *Aspergillus* and *Exophiala* species including two tyrosinase enzymes, polyketide synthase, and alpha-beta hydrolase (Z. Chen et al., 2014). Finally, *Aspergillus* genomes are larger, 34 Mbp (*A. niger*), 30 Mbp (*A. nidulans*), 37.8 Mbp (*A. oryzae*), and 29.4 Mbp (*A. terreus*) ("National Center for Biotechnology Infornation," n.d.), than the *E. dermatitidis* genome, 26.4 Mbp ("National Center for Biotechnology Infornation," n.d.). The larger genome of *Aspergillus* species would likely encode for more metabolic functionalities that *E. dermatitidis*. However, model Ascomycota species likely encode for fewer metabolic functionalities in that their genomes are significantly smaller than *E. dermatitidis*, 11.8 Mbp for *S. cerevisiae* and 20.2 Mbp for *Y. lipolitica*. For these reasons, it was deemed appropriate to restrict the databases associated with filling metabolic gaps to functionalities identifiable in *Aspergillus* species.

3.6.3. Consensus of Aspergillus models

The relatively small percentages of ORFs with assigned EC numbers by annotations or by information from BRENDA suggested the need to explore GSMs of related species to identify core enzymes. The *Aspergillus* genus was identified as closely related (Schoch et al., 2009) with four GSM models: *A. niger* (Andersen et al., 2008)*, A. nidulans* (David et al., 2008)*, A. oryzae* (Vongsangnak et al., 2008)*,* and *A. terreus* (J. Liu et al., 2013). Models of *Saccharomyces cerevisiae* were not used at this stage of the curation process as *S. cerevisiae* is phylogenetically much more distant to *E. dermatitidis* than are *Aspergillus* species (Schoch et al., 2009). Using the GPR links in the *Aspergillus* models, provided in the form of EC numbers, the links were sorted into bins of full consensus, common to three, common to two, and only present in one. EC numbers in the full consensus bin were added to the *Exophiala* model by repeating the process used on the *Exophiala* EC numbers and using the compartmentalization from the *Aspergillus* models. The other bins were similarly converted to lists of reactions with compartmentalization from the *Aspergillus* models. These three reaction lists are the databases used in the application of OptFill (Schroeder & Saha, 2020b) to the *Exophiala* model. Further, transport reactions were taken from the *Aspergillus* models and added to the *Exophiala* model as needed.

3.6.4. Bidirectional BLASTp of Aspergillus consensus enzymes

In order to create GPR links between *Aspergillus* enzymes and reactions added to the *Exophiala* models, a program which performs a bidirectional BLASTp on a list of enzymes subject to certain constraints and with the intent of identifying a gene encoding each enzyme in the list was created, and is included in GitHub "E_dermatitidis_model" repository. The constraint file specifies the target organism (in this work, *Exophiala*), the file path in which FASTAs and BLASTp results will be deposited to, the expect value upper bound cut-off (for which a match to be accepted), the percent positive substitution lower bound value (for matches in which the percent positive substitution value being greater than or equal to the cut off), and related species from which to take reference sequences for an enzyme. The Bidirectional Blast Program (hereafter BBP) begins by reading the constraints and enzyme list files. The workflow followed by the BBP is shown in Supplemental Figure SDD. In short, BBP takes the enzyme classification (EC) number and uses it to look up the amino acid sequences for genes encoding that EC number in the related organism. The amino acid sequences (in the form of a FASTA file) of the related organism (in this work, *A. nidulans*, *A. niger*, *A. terreus* or *A. oryzae*) is BLASTed against the target organism (in this work, *Exophiala*). This is called the “forward BLAST”. Should this forward BLAST result in a significant match according to the cutoffs specified in the constraints file, the amino acid sequence from *Exophiala* is then BLASTed against the related organism from which the EC producing amino acid sequence is taken. This is referred to as the “backward BLAST”. Should the backward BLAST results be significant according to the cutoffs in the constraint file, the match between the two sequences is accepted. A summary of the BLAST results can be found in Supplemental File 2, FASTAs for the sequences can be found in the GitHub “E_dermatitidis_model” repository.

3.6.5. Definition of biomass composition

The definition of biomass is provided in supplemental File 4. Literature evidence was sought out for the biomass composition of *Exophiala*, beginning with the composition of the cell wall. The cell wall composition can be found in Table 2 and is from data for *Exophiala* grown at $37^{o}\ C$ (Philip Anthony Geis, 1981). As this work did not distinguish between types of melanin in the cell wall of *Exophiala*, and no follow-up study was found which made this distinction, each of the three melanins, namely DNH melanin, eumelanin, and pyomelanin, is assumed to contribute equally to the cell wall mass. Further, the composition of the lipid term was unspecified. No data was found as to the lipid composition of *Exophiala*, and therefore the lipid composition of *A. terreus* grown using glucose as a carbon source at $28^{0}\ C$ was used (A. K. Kumar & Vatsyayan, 2010). Only lipids with KEGG identifiers were included in the lipid composition definition, accounting for 82.3% of the lipid composition of *A. terreus* (by weight percentage) (A. K. Kumar & Vatsyayan, 2010). In addition, there was no information on the fraction of *Exophiala* cell mass that was accounted for by the cell wall itself; however, *S. cerevisiae* cell walls account for 15-30% of the total cell mass (Lipke & Ovalle, 1998). For *Exophiala* , it was assumed that 25% of the cell mass is cell wall as the cell wall of this species has been described as “thick” (Z. Chen et al., 2014)(A.

K. Kumar & Vatsyayan, 2010)(Schnitzler et al., 1999). The next set of literature evidence sought for biomass composition is the cell without the cell wall. No literature evidence was found, and so to be consistent with the lipid composition information and considering that the *A. terreus* model is the most recently published of the four *Aspergillus* models, its biomass composition was used for the cell. Finally, the carotenoid contribution to biomass was considered. Again, lack of information pertaining to *Exophiala* led to using carotenoid biomass composition data from another organism, in this case, *Podospora anserina*, for which approximately 3.47% of cell mass is carotenoids (Strobel et al., 2009). Both *P. anserina* and *Exophiala* are Ascomycota, but phylogenetically diverge at the class level. Unfortunately, no similar data was able to be found for a species which was phylogenetically closer to *Exophiala*. The remaining cell weight, 71.53%, was assumed to be composed of the cell membrane and all biomass components enclosed within it (such as proteins and lipids). As the ratios of carotenoids, cell wall, and other cell components were determined to be important, biomass composition was divided into cell wall, cell, and carotenoid pseudometabolites. These pseudometabolites represented the mass contributions of fixed stoichiometeries of metabolites which comprise that portion of biomass. Each of these then had their own pseudo-molecular weight. These three pseudometabolites were then combined in a pseudoreaction to form biomass. The stoichiometric ratio of these pseudometabolites in the biomass reaction was determined by using the Solver tool in Microsoft Excel whose objective was a biomass molecular weight of 1000 mg/gDW·h. Ratios between pseudometabolites were enforced in this analysis to preserve ratios as described above. The resulting biomass composition can be found in Supplemental File 4.

3.6.6. Creation of first and second drafts of E. dermatitidis GSM

The first draft of the *Exophiala* GSM was the combination of the set of reactions which exist in *Exophiala* as determined by the analysis of the annotated genomes, the analysis of consensus *Aspergillus* enzymes, and the defined biomass composition. This first draft model did not produce biomass, melanins, or carotenoids. Through manual curation and addition of reactions related to melanin and carotenoid synthesis, both classes of pigments were produced by the second *Exophiala* draft model. Thermodynamically Infeasible Cycle (TICs) in the draft model were manually addressed. Further, some reactions were needed to be manually added to the model to ensure biomass production. These reactions were taken from a GSM of *S. cerevisiae*, *i*Sce926 (Chowdhury et al., 2015), and a list of reactions derived from the common to four *Aspergillus* models enzyme overlap (Andersen et al., 2008)(David et al., 2008)(Vongsangnak et al., 2008)(J. Liu et al., 2013) using code included in the GitHub "E_dermatitidis_model" repository. Reactions from the common to three *Aspergillus* list which were used in manual curation were removed from that list before performing OptFill on the second draft model. Notes related to the curation process can be found in the GitHub "E_dermatitidis_model" repository. Once TICs were eliminated, the model could produce all pigment molecules and biomass and utilize multiple literature-supported carbon sources including sucrose, ethanol, acetate, and glucose. The second draft *Exophiala* model was then considered complete. This second draft model still had a significant number of metabolic gaps, particularly in secondary metabolism, as evidenced by Flux Variability Analysis (FVA) (Steinn Gudmundsson & Thiele, 2010). When applied to this model, FVA showed that only 711 of 1587 reaction present in the model were capable of carrying flux (about 44.8%). This model could produce 591 metabolites (of a total of 1839). The maximum rate of growth of this model was 0.0952 h^{-1}.

3.6.7. Update to the OptFill algorithm

In the process of applying OptFill to draft models of *Exophiala*, which are the largest database/model pairs to which OptFill has thus far been applied (Steinn Gudmundsson & Thiele, 2010), it was discovered that additional constraints were necessary in order that the algorithm is not sensitive to solver options used. Specifically, these constraints were required in the Connecting Problems (CPs) of OptFill, and are listed below in equations (3.13) through (3.17), along with the full formulation for the first CP. The second and third CPs used here are related to the first CP in the same manner as detailed in Schroeder and Saha (2020).

$$\boldsymbol{maximize\ Z_{met} = \sum_{i_m \in I^M} x_{i_m}} \tag{3.1}$$

Subject to

$$\boldsymbol{\sum_{j_{db} \in J^{DB}} \zeta_{j_{db}} \geq 1} \tag{3.2}$$

$$\boldsymbol{\rho_{j_{db}} v_{j_{db}}^{LB} \leq v_{j_{db}} \leq \delta_{j_{db}} v_{j_{db}}^{UB}} \qquad \forall \boldsymbol{j_{db} \in J^{DB}} \tag{3.3}$$

$$\boldsymbol{\theta_j v_j^{LB} \leq v_j \leq (1 - \theta_j) v_j^{UB} - \epsilon \theta_j} \qquad \forall \boldsymbol{j \in J} \tag{3.4}$$

$$\boldsymbol{(1 - \lambda_j) v_j^{LB} + \epsilon \lambda_j \leq v_j \leq \lambda_j v_j^{UB}} \tag{3.5}$$

$$\boldsymbol{x_i \leq \sum_{j \in J} [\lambda_j \xi_{i,i} + \theta_j \psi_{i,j}]} \qquad \forall \boldsymbol{i \in I} \tag{3.6}$$

$$\boldsymbol{x_b = 1} \qquad \forall \boldsymbol{b \in B \subset I} \tag{3.7}$$

$$\boldsymbol{\sum_{j \in J} S_{ij} v_j = 0} \qquad \forall \boldsymbol{i \in I} \tag{3.8}$$

$$\boldsymbol{\zeta_{j_{db}} \leq \sum_{i \in I} [\lambda_j \xi_{i,i} + \theta_j \psi_{i,j}]} \qquad \forall \boldsymbol{j_{db} \in J^{DB}} \tag{3.9}$$

$$\boldsymbol{\delta_{j_{db}} + \rho_{j_{db}} - \omega_{j_{db}} = \zeta_{j_{db}}} \qquad \forall \boldsymbol{j_{db} \in J^{DB}} \tag{3.10}$$

$$\boldsymbol{\omega_{j_{db}} \leq \delta_{j_{db}}} \qquad \forall \boldsymbol{j_{db} \in J^{DB}} \tag{3.11}$$

$$\boldsymbol{\omega_{j_{db}} \leq \rho_{j_{db}}} \qquad \forall \boldsymbol{j_{db} \in J^{DB}} \tag{3.12}$$

$$\boldsymbol{\omega_{j_{db}} \leq \theta_{j_{db}} + \lambda_{j_{db}}} \qquad \forall \boldsymbol{j_{db} \in J^{DB}} \tag{3.13}$$

$$\boldsymbol{\rho_{j_{db}} \leq \theta_{j_{db}} + \lambda_{j_{db}}} \qquad \forall \boldsymbol{j_{db} \in J^{DB}} \tag{3.14}$$

$$\boldsymbol{\delta_{j_{db}} \leq \theta_{j_{db}} + \lambda_{j_{db}}} \qquad \forall \boldsymbol{j_{db} \in J^{DB}} \tag{3.15}$$

$$\boldsymbol{\delta_{j_{db}} \geq v_{j_{db}}} \qquad \forall \boldsymbol{j_{db} \in J^{DB}} \tag{3.16}$$

$$\boldsymbol{\rho_{j_{db}} \geq v_{j_{db}}} \qquad \forall \boldsymbol{j_{db} \in J^{DB}} \tag{3.17}$$

$$\boldsymbol{\sum_{j_{db} \in J^{DB}} \delta_{j_{db}} \left(\delta'_{s_c,j_{db}}\right) \leq \sum_{j_{db} \in J^{DB}} \delta'_{s_c,j_{db}} - \sigma_{s_c}} \qquad \forall \boldsymbol{s_c \in S_c} \tag{3.18}$$

$$\boldsymbol{\sum_{j_{db} \in J^{DB}} \rho_{j_{db}} \left(\rho'_{s_c,j_{db}}\right) \leq \sum_{j_{db} \in J^{DB}} \rho'_{s_c,j_{db}} - \left(1 - \sigma_{s_c}\right)} \qquad \forall \boldsymbol{s_c \in S_c} \tag{3.19}$$

$$\boldsymbol{\sum_{j_{db} \in J^{DB}} \left(\delta'_{s_c,j_{db}} - \delta_{j_{db}}\right) + \sum_{j_{db} \in J^{DB}} \left(\rho'_{s_c,j_{db}} - \rho_{j_{db}}\right) \geq \left(\sum_{j_{db} \in J^{DB}} \omega'_{s_c,j_{db}}\right) + 1} \qquad \forall \boldsymbol{s_c \in S_c} \tag{3.20}$$

$$\boldsymbol{\sum_{j_{db} \in J^{DB}} \delta_{j_{db}} \left(\alpha'_{s_f,j_{db}}\right) \leq \sum_{j_{db} \in J^{DB}} \alpha'_{s_f,j_{db}} - \tau_{s_f}} \qquad \forall \boldsymbol{s_f \in S_f} \tag{3.21}$$

$$\boldsymbol{\sum_{j_{db} \in J^{DB}} \rho_{j_{db}} \left(\beta'_{s_f,j_{db}}\right) \leq \sum_{j_{db} \in J^{DB}} \beta'_{s_f,j} - \left(1 - \tau_{s_f}\right)} \qquad \forall \boldsymbol{s_f \in S_f} \tag{3.22}$$

Where symbols used are defined as follows.

$\boldsymbol{Fixed\ Values\ Unique\ to\ CP1}$

$\boldsymbol{M = 1E3 \equiv a\ very\ large\ number}$

$$\boldsymbol{\delta'_{s_c,j_{db}} = \begin{cases} 1\ if\ reaction\ j_{db}\ is\ added\ in\ the\ forward\ direction\ from\ the \\ \qquad database\ in\ solution\ s_c \\ 0\ otherwise \end{cases}}$$

$$\boldsymbol{\rho'_{s_c,j_{db}} = \begin{cases} 1\ if\ reaction\ j_{db}\ is\ added\ in\ the\ backward\ direction\ from\ the \\ \qquad database\ in\ solution\ s_c \\ 0\ otherwise \end{cases}}$$

$$\boldsymbol{\omega'_{s_o,j_{db}} = \begin{cases} 1\ if\ reaction\ j_{db}\ is\ added\ in\ the\ forward\ direction\ from\ the \\ \qquad database\ in\ solution\ s_o \\ 0\ otherwise \end{cases}}$$

$$\xi_{i,j} = \begin{cases} 1 \ if\ metabolite\ i\ is\ on\ the\ RHS\ of\ reaction\ j\ \left(S_{i,j} > 0\right) \\ 0\ otherwise \end{cases}$$

$$\psi_{i,j} = \begin{cases} 1 \ if\ metabolite\ i\ is\ on\ the\ LHS\ of\ reaction\ j\ \left(S_{i,j} < 0\right) \\ 0\ otherwise \end{cases}$$

$Variables\ Unique\ to\ CP1$

$$\delta_{j_{db}} = \begin{cases} 1\ if\ reaction\ j_{db}\ is\ added\ in\ the\ forward\ direction\ from\ the\ database \\ 0\ otherwise \end{cases}$$

$$\rho_{j_{db}}$$

$$= \begin{cases} 1\ if\ reaction\ j_{db}\ is\ added\ in\ the\ backwards\ direction\ from\ the\ database \\ 0\ otherwise \end{cases}$$

$$\omega_{j_{db}}$$

$$= \begin{cases} 1\ if\ reaction\ j_{db}\ is\ added\ reversibly\ from\ the\ database\ (\delta_{j_{db}} = \rho_{j_{db}} = 1) \\ 0\ otherwise \end{cases}$$

$$\zeta_{j_{db}} = \begin{cases} 1\ if\ reaction\ j_{db}\ is\ part\ of\ the\ solution \\ 0\ otherwise \end{cases}$$

$$\theta_j = \begin{cases} 1\ if\ reaction\ is\ proceding\ in\ backwards\ direction\ \left(v_j < 0\right) \\ 0\ otherwise \end{cases}$$

$$\lambda_j = \begin{cases} 1\ if\ reaction\ is\ proceding\ in\ forwards\ direction\ \left(v_j > 0\right) \\ 0\ otherwise \end{cases}$$

$$x_i = \begin{cases} 1\ if\ connected\ model\ produces\ metabolite\ i \\ 0\ otherwise \end{cases}$$

$\sigma_{s_c} \in [0,1] \equiv binary\ variable\ which\ ensures\ that\ the\ solution\ is\ unique\ from$

$previous\ solutions\ in\ at\ least\ one\ direction\ of\ one\ database\ reaction$

$\tau_{s_c} \in [0,1]$

$\equiv binary\ variable\ which\ ensures\ that\ the\ solution\ is\ free\ from\ TICs\ in$

$that\ at\ least\ one\ direction\ of\ one\ database\ reactions\ which\ could\ cause\ a\ TIC\ is$

$added\ in\ the\ TIC - causing\ direction\ for\ each\ TIC\ identified\ by\ the\ TFP$

Equations displayed above, with the exception of equations (3.13) through (3.17), are identical to the original formulation of OptFill. In short, additional constraints (3.13), (3.14), and (3.15) explicitly link binary variables noting the reaction direction and binary variables relating the direction of database reactions which are added to the model. This reduces the impact of any feasibility relaxation assumptions made by the solver in attempting to solve the CPs. Additional constraints (3.16) and (3.17) restrict the range of each reaction rate, v_j, to be less than one (as flux magnitude is not important and this allows effectively tighter optimization criteria), while at the same time requiring $\delta_{j_{db}}$ and $\rho_{j_{db}}$ to have a non-zero value if the metabolic flux through that database reaction is non-zero. Theoretically, this addressed by constraint equation (3.3), but these statements again reduce the effect of feasibility relaxations.

3.6.8. First use of OptFill to address metabolic gaps

To address the metabolic gaps in the second draft model, OptFill (Steinn Gudmundsson & Thiele, 2010) was used first with a list of reactions derived from the list of enzymes common to four *Aspergillus* models. The database was reduced to a manageable size (e.g. one that would allow at least one solution to the CPs in less than one week) over six rounds of identifying TICs and pruning the database of reactions which caused the most TICs. By the end of this pruning, the database consisted of 241 reactions, had 82 potential TICs with the model (largest size 12 reactions), and two connecting problem solutions. The first solution, which could produce 620 metabolites by adding 20 reversible reactions, was accepted over the second solution, which could produce 619 metabolites by adding seven reversible and 11 irreversible reactions. Adding the first CPs solution to the second draft model produced the third draft model consisting of the 1607 reactions, of which 749 are capable of carrying flux (about 46.6%). The maximum rate of growth of this model was 0.0989 h^{-1} by allowing for up to 10 mmol·gDW^{-1}·h^{-1} uptake of one of ethanol, sucrose, glucose, and acetate along with sufficient amount of nitrate, sulfate, and phosphate.

3.6.9. Second and third use of OptFill to address metabolic gaps

This process was repeated twice: the first time using the list of reactions derived from the list of enzymes common to two *Aspergillus* models, and the second time using the list of reactions derived from the list of enzymes common to two *Aspergillus* models. For more details on the results of each step, see Supplemental File 5. The end result is the final model, *i*Ede2091, consisting of 1630 reactions, of which 793 are capable of carrying flux (48.7%). The maximum rate of growth of this model was 0.0989 h^{-1} by allowing for 10 mmol·gDW^{-1}·h^{-1} uptake of ethanol, sucrose, glucose, acetate, nitrate, sulfate, and phosphate. In this growth condition, carbon is the limiting nutrient. It should be noted that, on a minimal media where sucrose is provided as the primary carbon source, that is at a concentration two orders of magnitude higher than any other potential carbon source, the growth rate of *Exophiala* is approximately 0.105 h^{-1} (Dadachova et al., 2007); however, since no rate measures were taken in the indicated study, it is difficult to interpret the accuracy of the modeled growth rate of *i*Ede2091.

3.6.10. Bidirectional BLASTp to investigate OptFill solution viability

For each OptFill solution incorporated into the draft models, a bidirectional BLASTp analysis was performed on enzymes linked to the reactions in each OptFill solution. For the first OptFill solution, containing reactions linked to enzymes common to three of four *Aspergillus* models, 20 enzymes were identified as linked to this set of reactions. Using the same bidirectional BLASTp procedure as previously described, 11 of these enzymes were identified in the *E.*

dermatitidis genome, being matched to 21 genes. These genes were all annotated in the *E. dermatitidis* genome; therefore these enzymes may not have been identified by the BRENDA search of *E. dermatitidis* enzyme annotations due to sensitivity of the algorithm used for this search. These matches give a genetic basis for the inclusion of 11 of these reactions, in addition to the evidence that all these enzymes are supported in phylogenetically related organisms.

For the second OptFill solution, containing 3 reactions, 3 enzymes were identified as linked to the set of reactions in the solution, and two of these enzymes where identified in the *E. dermatitidis* genome. These two enzymes were linked to two genes. These genes were all annotated in the *E. dermatitidis* genome; therefore these enzymes may not have been identified by the BRENDA search of *E. dermatitidis* enzyme annotations due to sensitivity of the algorithm used for this search to the annotated string. These matches give a genetic basis for the inclusion of 2 of these reactions, in addition to the evidence that all these enzymes are supported in phylogenetically related organisms.

For the third OptFill solution, containing 21 reactions, 17 enzymes were identified as linked to the set of reactions in the solution, and 8 of these enzymes where identified in the *E. dermatitidis* genome. These 8 enzymes were linked to 18 genes. These genes were all annotated in the *E. dermatitidis* genome; therefore these enzymes may not have been identified by the BRENDA search of *E. dermatitidis* enzyme annotations due to sensitivity of the algorithm used for this search to the annotated string. These matches give a genetic basis for the inclusion of 13 of these reactions, in addition to the evidence that all these enzymes are supported in phylogenetically related organisms.

3.6.11. Flux Balance and shadow price analyses

Flux Balance Analysis (FBA) is a tool to study distribution of fluxes subject to an objective function (often growth) and certain constraints (e.g., mass balance and nutrient availability) for an underdetermined network (Orth et al., 2010), and was performed as previously described (Orth et al., 2010)(Gianchandani, Chavali, & Papin, 2010). The dual formulation of FBA, and the definition of shadow price, was derived as described by Zomorrodi and Costas (Maranas & Zomorrodi, 2016). The shadow price of a metabolite is the change in the value of the objective function used in FBA (growth) that would result from producing one more unit (mmol·gDW^{-1}·h^{-1}) of that metabolite. The shadow price is the λ_i variable of the dual formulation of the FBA problem shown below. The shadow price relating to all 36 growth conditions was calculated using the primary and dual formulations shown below. The primal FBA problem is as follows.

$$\boldsymbol{maximize\ z_{prime} = v_{biomass}} \tag{23}$$

Subject to:

$$\sum_{j \in J} \boldsymbol{S_{ij} v_j = 0} \qquad \forall \boldsymbol{i} \in \boldsymbol{I} \tag{24}$$

$$\boldsymbol{v_j^{LB} \leq v_j \leq v_j^{UB}} \qquad \forall \boldsymbol{j} \in \boldsymbol{J} \tag{25}$$

The dual FBA problem is as follows.

$$\boldsymbol{maximize\ z_{dual}} = \sum_{j \in J} \boldsymbol{v_j^{UB} \mu_j^{UB}} + \sum_{j \in J} \boldsymbol{v_j^{LB} \mu_j^{LB}} \tag{26}$$

Subject to:

$$\sum_{i \in I} \boldsymbol{S_{ij} \lambda_i - \mu_j^{LB} + \mu_j^{UB} = 0} \qquad \forall \boldsymbol{j} \in \boldsymbol{J - biomass} \tag{27}$$

$$\sum_{i \in I} \boldsymbol{S_{ij} \lambda_i - \mu_j^{LB} + \mu_j^{UB} = 1} \qquad \forall \boldsymbol{j} \in \boldsymbol{biomass} \tag{28}$$

By applying strong duality theory using the following constraint, both primal and dual variables may be explicitly solved.

$$\boldsymbol{z_{dual}} = \boldsymbol{z_{primal}} \tag{29}$$

In the reconstruction of the *i*Ede2091 model, it was noted that the availability of three nutrient atoms, namely carbon, nitrogen, and sulfur, could limit the growth of the *i*Ede2091 model. Further the model can grow on four different carbon sources. Tt was decided to investigate the effect of growth limiting nutrients on the shadow price of defensive pigments and their precursors under 36 unique growth conditions, where each carbon source/limiting atom pair is investigated under low, moderate, and high availability.

3.6.12. Identification of E. dermatitidis tyrosinase enzymes and attempted identification of tyrosinase related proteins.

Four tyrosinase enzymes have been annotated in the *Exophiala* genome and noted in literature (Z. Chen et al., 2014). In order to ensure that all gene copies of *Exophiala* tyrosinase were accounted for, the four tyrosinase genes from *E. dermatitidis* were BLASTed against the *Exophiala* genome using non-redundant BLASTp. No accessions which were not previously annotated as tyrosinase were identified, see the GitHub “E_dermatitidis_model” repository for the BLAST results related to tyrosinase. As tyrosinase-related proteins share high sequence similarity to tyrosinases of a species (Furumura et al., 1998), all four tyrosinase sequences for *Exophiala* were BLASTed against its own genome, again using non-redundant BLASTp. No significant matches were found except for known tyrosinases. As there is no literature evidence for *Exophiala* or

Aspergillus species with tyrosinase-related proteins, it was concluded from this that no tyrosinase-related proteins are encoded for by the *Exophiala* genome.

3.6.13. Comparison of E. dermatitidis tyrosinase gene copies to Hidden Markov Model (HMM) tyrosinase sequences.

A comparison of each tyrosinase sequence in *E. dermatitidis* was made to the Hidden Markov Model tyrosinase sequence using the Pfam tool (El-Gebali et al., 2019). The amino acid sequence for each *E. dermatitidis* gene copy was used as the search sequence. All gene copies had strong sequence alignments to the tyrosinase HMM, with gene copies unique to *Exophiala* matching weakly to the tyrosinase C HMM as well. All sequence alignments had expect values between 2.0E-38 and 3.1E-54, showing very strong agreement.

3.6.14. Comparison of E. dermatitidis tyrosinase gene copies to human tyrosinase alleles.

First, the amino acid sequences of *Exophiala* tyrosinases were BLASTed against the human genome using non-redundant BLASTp. This produced no significant matches (see the tyrosinase sequence alignments provided in the GitHub "E_dermatitidis_model" repository) initially suggesting that these enzymes were quite different. However, a COBALT amino acid sequence alignment was performed comparing three human alleles for tyrosinase, tyrosinase-related protein sequences from human, and the tyrosinase reference sequences for *Exophiala*. For the human alleles chosen, one was a reference sequence, one an albino sequence for oculocutaneous albinism A1, and one sequence from an individual of the Bantu peoples of Kenya (Hudjashov, Villems, & Kivisild, 2013), representing a population susceptible to the ill-effects of albinism (Brilliant, 2015). An independent COBALT sequence alignment was also performed with only the three human alleles selected to identify the sequential differences between the three alleles. Both COBALT alignments are provided in the GitHub "E_dermatitidis_model" repository, and a visualization of the results is provided in Figure 3.5, with particular attention paid to the active site of tyrosinase which is the two copper-binding domains. Visualization highlighting uses the 3-bit conservation score setting was used for highlighting sequence similarities as it seems a moderately-strict setting and no standard for this highlighting scheme was identified in literature. Literature was used to identify both tyrosinase active sites, CuA, from approximately residues 173 to 220 in human tyrosinase (Furumura et al., 1998)(García-Borrón & Solano, 2002), and CuB, from approximately residues 361 to 403 in human tyrosinase (Furumura et al., 1998)(García-Borrón & Solano, 2002)(Spritz, Ho, Furumura, & Hearing, 1997). Labels for the significance of highly conserved residues are taken from the analysis of García-Borrón and Solano (2002) (García-Borrón & Solano, 2002).

Chapter 4

4. AN OPTIMIZATION- AND EXPLICIT RUNGE-KUTTA- BASED APPROACH TO PERFORM DYNAMICA FLUX BALANCE ANALYSIS

Portions of this material have previously appeared in the following publication:

W. L. Schroeder, S. D. Harris, and R. Saha, Computation-Driven Analysis of Model Polyextremotolerant Fungus Exophiala dermatitidis: Defensive Pigment Metabolic Costs and Human Applications, iScience, 23(2020) 1-17. Used with permission.

W. L. Schroeder, R. Saha, Protocol for Genome-Scale Reconstruction and Melanogenesis Analysis of Exophiala dermatitidis, STAR Protocols, 1(2020) 1-37. Used with permission.

4.1. PREFACE

In this chapter we introduce the generalized optimization- and explicit Runge-Kutta-based Approach (ORKA) to perform dynamic flux Balance Analysis (dFBA), which is numerically more accurate and computationally tractable than existing approaches. ORKA is applied to a four-tissue (leaf, root, seed, and stem) model of Arabidopsis thaliana, p-ath773, uniquely capturing the core-metabolism of several stages of growth from seedling to senescence at hourly intervals. Model p-ath773 has been designed to show broad agreement with published plant-scale properties such as mass, maintenance, and senescence, yet leaving reaction-level behavior unconstrained. Hence, it serves as a framework to study the reaction-level behavior necessary for observed plant-scale behavior. Two such case studies of reaction-level behavior include the lifecycle progression of sulfur metabolism and the diurnal flow of water throughout the plant. Specifically, p-ath773 shows how transpiration drives water flow through the plant and how water produced by leaf tissue metabolism may contribute significantly to transpired water. Investigation of sulfur metabolism elucidates frequent cross-compartment exchange of a standing pool of amino acids which is used to regulate the proton flow. Overall, p-ath773 and ORKA serve as scaffolds for dFBA-based lifecycle modeling of plants and other systems to further broaden the scope of in silico metabolic investigation.

4.2. INTRODUCTION

In addition, tools which expand on the functionality of the basic FBA formulation, such as dynamic FBA (dFBA) (Mahadevan, Edwards, & Doyle, 2002) can improve the predictive abilities of SMs. dFBA can perform FBA over windows of time by solving a dynamic non-linear or a static linear problem, both of which integrate system variables over discrete time windows to solve for metabolite concentrations, reaction fluxes, and system biomass (Mahadevan et al., 2002; Grafahrend-Belau et al., 2013). In general, there are two approaches to dFBA. First, the Static Optimization-based Approach (SOA) which has been applied to *E. coli* (Mahadevan et al., 2002), mammalian cells (Luo et al., 2006; Bordbar et al., 2017), *Saccharomyces cerevisiae* (bakers' yeast)[18], *Hordeum vulgare* (barley) (Grafahrend-Belau, Schreiber, Koschutzki, & Junker, 2009), and *Arabidopsis thaliana* (Shaw & Cheung, 2018) (in addition to other systems). Second, the Dynamic Optimization-based Approach (DOA) which has been applied to *E. coli* metabolism (Grafahrend-Belau et al., 2009) and signaling networks is *S. cerevisiae* (Min Lee, Gianchandani, Eddy, & Papin, 2008) (to name a few applications). These approaches have proven useful for investigating aspects of plant-scale metabolism, such as resource partitioning in *Arabidopsis* (Shaw & Cheung, 2018). These works have inspired the development of our new approach to perform dFBA named as Optimization- and explicit Runge-Kutta –based Approach (ORKA). ORKA significantly improves upon the SOA by utilizing the step-by-step solution approach of the SOA (as opposed to simultaneous solution of all times in the DOA) with increased accuracy and solution stability. These improved characteristics are due to both the implementation of a Runge-Kutta method (a multi-step numerical method for the solution of ordinary differential equations) to replace the first-order Taylor series approximation used by SOA and by replacing the assumption that the reaction rate is constant over each time interval with a trapezoid rule-based integral approximation.

Arabidopsis thaliana (hereafter *Arabidopsis*) has been selected as a test system for the application and demonstration of the ORKA framework, due to the fact that *Arabidopsis* is a model plant species with a highly characterized knowledgebase. The choice would also allow demonstration of ORKA in a dynamic, multi-tissue system. To date, many stoichiometric models of plant metabolism, including *Arabidopsis*, have been developed. Some of these models including models of *Arabidopsis thaliana* (Gomes de Oliveira Dal'Molin et al., 2015; Grafahrend-Belau et al., 2013; Poolman, Miguet, Sweetlove, & Fell, 2009; de Oliveira Dal'Molin et al., 2010), *Zea mayz* (maize) (Saha et al., 2011), *Sorghum bicolor* (sorghum) (de Oliveira Dal'Molin et al., 2010), *Brassica napus* (rapeseed) (Pilalis, Chatziioannou, Thomasset, & Kolisis, 2011), and *Oryza sativa* (rice) (M. G. Poolman, Kundu, Shaw, & Fell, 2013) have treated plants as single metabolic units. These models have sought to analyze metabolic maintenance, response to abiotic stimuli, enzyme regulation changes, and metabolism as a whole (de Oliveira Dal'Molin et al., 2010; Gomes de Oliveira Dal'Molin et al., 2015; Grafahrend-Belau et al., 2013; Poolman et al., 2009; de Oliveira Dal'Molin et al., 2010; Saha et al., 2011; Pilalis, Chatziioannou, Thomasset, & Kolisis, 2011; M. G. Poolman et al., 2013). Tissue-specific models have been reconstructed for various *Arabidopsis* tissues (Mintz-Oron et al., 2012), a maize leaf (Simons et al., 2014), and a barley seed (Mahadevan et al., 2002) to better understand how present metabolites, metabolic pathways, and nutrient (generally carbon and nitrogen) availability differ between tissues. Multi-tissue models have also been developed to characterize whole-plant metabolism for *Arabidopsis* (Gomes de Oliveira Dal'Molin et al., 2015; Shaw & Cheung, 2018) and barley (Luo et al., 2006) and subsequently to study whole-plant metabolic response to the diurnal cycle and the source-to-sink relationship of leaves and seeds (Grafahrend-Belau et al., 2009; (Gomes de Oliveira Dal'Molin et al., 2015). These studies have considered metabolism at a single point (often in the exponential growth phase (de Oliveira Dal'Molin et al., 2010; Grafahrend-Belau et al., 2009; Poolman et al., 2009; de Oliveira Dal'Molin et al., 2010; Saha et al., 2011; M. G. Poolman et al., 2013)) or a single diurnal cycle (Gomes de Oliveira Dal'Molin et al., 2015) or have modeled only a portion of the *Arabidopsis* lifecycle (Shaw & Cheung, 2018). The most complete dFBA work, in terms of modeling the full *Arabidopsis* lifecycle, models two tissues, leaf and root, across 30 days of vegetative growth (from 6 days to 36 days) (Shaw & Cheung, 2018). Here, we have developed a core-carbon metabolic model of *Arabidopsis*, named p-ath773 (plant-scale core-metabolism *Arabidopsis thaliana* model with 773 genes included), to model the full lifecycle of *Arabidopsis* from germination to senescence by being embedded in the ORKA framework which captures metabolic interactions between four major tissues: leaf, root, seed, and stem. These four tissues have been chosen for model reconstruction to represent core plant functions: the root for nutrient uptake and growth; the leaf for photosynthesis, carbon fixation, and as a source tissue for plant nutrition; the seed for metabolite storage and a sink tissue for metabolic investment; and the stem for metabolic transport and acting as a conduit for all metabolic interactions between other tissues. Core-metabolism pathways that are included but not limited to photosynthesis; the citrate cycle; starch and sucrose synthesis; fatty acid synthesis and degradation; and amino acid synthesis. The p-ath773 model consists of 1251 total (and 631 unique, defined as having the same identifier across any number of subcellular compartments) reactions (R), 1155 total (and 276 unique) metabolites (M), and accounts for 773 genes (G) including 42 chloroplastic and 11 mitochondrial genes. Each of the modelled tissues including leaf (R: 517, M: 463, and G: 666), root (R: 149, M: 149, and G: 324), seed (R: 418, M: 390, and G: 577), and stem (R: 167, M: 154, and G: 291) has been reconstructed individually to allow for the different tissue mass ratios found across the lifecycle of the plant. A summary of the p-ath773 model is shown in Figure 4.1. The ORKA framework determines biomass, metabolite concentrations, reaction flux, change in biomass, and changes in metabolic concentration (collectively defined as a metabolic "snapshot") hourly across the lifecycle of *Arabidopsis* as modeled by p-ath773 under 12 hour light and 12 hour darkness growth conditions accounting for changes due to diurnal metabolic differences; changes in plant mass; metabolite storage and uptake (particularly carbohydrates); changes in plant tissue mass ratios; and changes in metabolism with respect to plant growth stage. The p-ath773 model is unique among *Arabidopsis* models for its focus on plant-scale behavior such as focus on achieving biomass levels which correspond with *in vivo* data; biomass-based maintenance and senescence drains; and the logical mole-balanced exchange of nutrients between tissues. While the plant-scale behavior is well-constrained in the p-ath773 model, reaction-scale behavior is unconstrained such that the model can be used to study the reaction-scale behavior necessary to explain observed macro-scale behavior.

When ORKA has been applied to the p-ath773 multi-tissue model, the order of error of both mass step and metabolite concentration estimates has been theoretically improved by approximately three order of magnitude as compared to that achieved in a previous model of *Arabidopsis* which utilized the SOA to perform dFBA (Shaw & Cheung, 2018). This has been done by combining improved mass step and metabolite concentration estimates with smaller time step sizes, one hour as opposed to one day (Shaw & Cheung, 2018). Further, with the inclusion of two more tissue types, stem and seed, and modeling the entire lifecycle, the p-ath773 model in the ORKA framework makes a significant improvement to current *Arabodipsis* dFBA-based models, despite only modelling central metabolism. It should be noted that for metabolic models with a single tissue, or single organism, O(h^3) or better error order is possible depending on the Runge-Kutta method selected, compared to the O(h^2) error order floor of the SOA method. This low error level has proved impossible to achieve with the p-ath773 model since the seed tissue appears and disappears over the course of the *Arabidopsis* lifecycle, causing difficulties due to the exponential nature of FBA-determined growth rates. The series of more accurate hourly metabolic "snapshots" produced by p-ath773 has given a framework for the investigation of the central metabolism of *Arabidopsis* across its lifecycle. Here, these "snapshots" have been used to investigate the diurnal patterns of water flow (from the root uptake to transpiration from the leaf), and sulfur metabolism (from root uptake to tissue biomass). Further, the p-ath773 model embedded in the ORKA framework has shown general agreement with macro-level experimental data found in the literature and is potentially useful as steppingstone for dynamic lifecycle modeling of other plant systems.

4.3. RESULTS

4.3.1. Development of the ORKA to perform dFBA

The Optimization- and explicit Runge-Kutta- based Approach (ORKA) has been developed to make more accurate and stable estimates of the changes in biomass and metabolite concentration in a dynamic Flux Balance Analysis (dFBA). The basis of the ORKA is the same as SOA, to model a dynamic (i.e. time-dependent) metabolism across multiple time points, where each time point solution builds upon previous solutions. The pseudocode describing how the ORKA works can be found in Figure 4.2A. Symbols are defined as follows: t_n is the current time, t_0 is the initial time, Δt is the time step, c_n are the steps in the independent variable made by the Runge-Kutta method chosen to use, Y_t is the current biomass concentration, Y_0 is the initial biomass concentration, a_{na} is the weight of Runge-Kutta derivative estimate steps (k_n) in the next derivative estimate, b_n is the weight of the Runge-Kutta derivative estimate steps in the full Runge-Kutta derivative estimate, $\left.\frac{dY}{dt}\right|_{rk\ est}$ is the Runge-Kutta derivative estimate for the current timestep, $Y_{t+\Delta t}$ is the biomass concentration at the next time step, $z_{i,t}$ is the concentration of metabolite i at time t, $z_{i,t+\Delta t}$ is the concentration of metabolite i at the next time step, S_{ij} is the stoichiometric coefficient of metabolite i in reaction j, $\Gamma_{j,t}$ is the trapezoid rule-based integral estimate of the flux of reaction

j at the current timestep, $v_{j,t}$ is the rate of reaction j at time t, v_{j,t_n} is the rate of reaction j at Runge-Kutta time step n, set N is the number of steps in the Runge-Kutta solution method (with n as the index), and c_{nf} is the final c_n value in the Runge-Kutta method. Greater detail on the definition of each symbols used can be found in the "Symbols Used" section (section 4.6.5). ORKA expands upon the SOA approach[28] by replacing the Taylor-series approximations (details in the methods section) used to advance biomass concentration in the SOA (Y_t in Figure 4.2A) with a Runge-Kutta-based estimate for increased model accuracy and solution stability. The ORKA framework in this pseudocode formulation is left generic enough so that a variety of Runge-Kutta methods can be used, as long as c_n values are evenly spaced. Here examples of Runge-Kutta methods which such c_n values include those shown in Butcher Tableaus in Figure 4.2. A detailed formulation of ORKA can be found in the Materials and Methods. A summary of the ORKA formulation is as follows. Begin with an SM; a set of time points over which to solve that SM; an initial condition related to the biomass of the system and metabolite concentrations; and a chosen Runge-Kutta method to use in the solution. For each time step, solve the SM using linear programming at the beginning of that time step and define the initial conditions (time, biomass, and metabolite concentrations). The chosen Runge-Kutta method is used to solve the change in those initial conditions over the time step. This is done by solving the SM using linear programming and saving all reaction rates for each solution for the given time step to determine the mass step estimate of the given Runge-Kutta step. Once all Runge-Kutta steps are complete, the final mass step estimate for the given time step is made. To advance metabolite concentration, the integral from the start of the time step to the end of the time step is estimated using the multi-application Trapezoid rule, which in turn is used to estimate the change in metabolite concentrations. This is followed by applying that mass step and concentration change estimates and repeating the process for the next time step. This process is shown more technically by a pseudocode described in Figure 4.2A and explained in full detail in Materials and Methods.

4.3.2. Reconstruction of Arabidopsis core metabolism in tissue-specific models

In order to track the important metabolic interactions and transactions within and between major tissues of *Arabidopsis* plant, namely seed, leaf, root, and stem, corresponding tissue-level metabolic models have been reconstructed. The seed and leaf tissue have been selected to model an important source-to-sink relationship, whereas the stem and root tissues have been included to model nutrient transport and nutrient uptake in *Arabidopsis*, respectively. Model files for each tissue can be found in the GitHub p-ath773 repository for this work (DOI: 10.5281/zenodo.3735103). Details of model reconstruction can be found in Materials and Methods, but a synopsis is as follows. The seed model has been reconstructed first using the metabolic pathways shown in the *Arabidopsis* seed though ^{13}C-labeled Metabolic Flux Analysis (MFA) (Lonien & Schwender, 2009). The model reactions have been distributed among extracellular space, cytosol, non-green plastid, inner mitochondria, and outer mitochondria subcellular compartments in accordance with literature evidence (see list of works cited in Data S1, see section 7.2 for how to access this file). Next, transport and exchange reactions have been added to the model based on literature evidence (see list of works cited in Data S1) or to increase model connectivity (Thiele & Palsson, 2010). The biomass composition of the seed has been determined from literature (Lonien & Schwender, 2009; Baud, Boutin, Miquel, Lepiniec, & Rochat, 2002). The resultant model is charge and element balanced, and has undergone multiple iterations of curation consistent with well-established GSM reconstruction protocols (Thiele & Palsson, 2010). Once the seed model has been reconstructed, metabolic pathways common to both the seed and leaf tissue have been used as the starting point for reconstructing the leaf tissue model. To this model have been added additional amino acid syntheses (for xylem and phloem loading), photosynthesis, and gluconeogenesis as well as chloroplast and thylakoid subcellular compartments. The biomass of the leaf has been adapted slightly from that of a previously published *Arabidopsis* model, *i*RS1597 (Saha et al., 2011), by refocusing the biomass composition on primary metabolites. Similarly, by having extracted common reactions/pathways from the seed and leaf models as a starting point and adding functionalities particular to these tissues such as nitrogen reduction in the root and the transport of metabolites through the extracellular space of the stem, the root and stem models have been reconstructed. Root and stem models have been reconstructed with metabolic differences between the two such as the presence of amino acid synthesis and the conversion of ammonium to nitrate both in the root for xylem loading. Root and stem tissues are, however, largely focus on basic carbon metabolism and metabolite uptake (root) and transport (root and stem). In the absence of *Arabidopsis*-specific estimates, the dry weight compositions of switchgrass (*Panicum virgatum*) root and stem (Baud et al., 2002) have been used to define root and stem biomass compositions. Both these models contain necessary transport/exchange reactions to ensure model connectivity and to facilitate their roles in the transport processes. The stem and root models have all the subcellular compartments present in the seed model. Once initial reconstructions have been accomplished, thermodynamically infeasible cycles in addition to atom and charge imbalances have been resolved (Thiele & Palsson, 2010). Figure 4.3 shows the iterative process of model curation for tissue-specific model reconstructions used in this work (yellow arrow) and for the whole-plant iterative model curation (orange arrows). Figure 4.4 shows a summary of the distribution of model reactions across KEGG-defined pathways of each tissue model and an overview of reasons for reaction inclusion through confidence scoring (see Method section) (Thiele & Palsson, 2010). Figure 4.4A summarizes the pathways common to all tissues and Figures 4.4B through 4.4E graphically summarize the sources of reactions in each tissue model through confidence scores (see methods section) (Thiele & Palsson, 2010). Once each tissue model has been reconstructed, these four models have been linked by the ORKA framework, and the lifecycle of the plant has been simulated. We have addressed the incongruities between these *in silico* simulation results and *in vivo* experimental data by adjusting their metabolism of individual tissue-specific models, tissue-tissue interactions, or by adjusting parameters (such as biomass yield, plant maintenance, and plant senescence) associated with the p-ath773 model. This portion of the workflow is illustrated in Figure 4.3 (orange arrows).

4.3.3. Development of constraints defining tissue-tissue interactions in the p-ath773 model

Once these core tissue models have been reconstructed and curated, sets of constraints have been defined to enforce logical links between tissues to facilitate the simulation of tissue metabolism. For instance, these links include ensuring that water travels from the root (source) to the leaves (sink) and that literature-supported amino acids travel from the leaf and root (sources) to the seed (sink) through the stem tissue (the link between these tissues). In addition, other constraints include environmental interactions such as with atmosphere and soil. These constraints include gas exchange in all tissues; uptake of micronutrients and water by the roots; and use of light by the leaves. These constraints are discussed in detail in the Materials and Methods. In summary, these constraints ensure that micronutrients and water are transported from the root tissue to other tissues via the stem; that sugars and amino acids travel from the leaf tissue to other tissues via the stem; that patterns of starch and sucrose storage in leaf and stem tissues are included in the model; and that the rates of tissue growth are linked in such a way that tissue mass ratios are preserved or changed in accordance with how these quantities change in an *Arabidopsis* plant as it passes through various stages of growth.

4.3.4. Simulating stages of plant growth using p-ath773 and ORKA

As discussed more extensively in Materials and Methods, the growth rate for an SM is an exponential growth rate. Due to this exponential nature of the growth rate, seed mass becomes problematic to model as there are points in the growth of the seed tissue where its mass is zero, is advanced from zero to a non-zero value, and is advanced from a non-zero value to zero. These conditions are impossible to capture using an exponential function. Therefore, plant mass as a whole is tracked and advanced by the ORKA, and individual tissues masses are determined by multiplying total plant mass by tissue mass fraction. Since there is no whole-plant biomass function, this approach requires an approximation which defines the error floor by a second order backward difference approximation of the first derivative (see Materials and Methods for details) with an error order of $O((c_2 - c_1)h^2)$. Therefore, any Runge-Kutta method with error order less than that will suffice. In this work, Heun's third-order Runge-Kutta rule is used. This is in part because of the limitation just described such that a higher-order Runge-Kutta method is not necessary. Further, this method has the advantage over Kutta's third-order Runge-Kutta rule in that the step size between c_n values is one third (e.g. $c_2 - c_1 = 1/3$) as opposed to one half (e.g. $c_2 - c_1 = 1/2$) (see Figure 4.2B), giving slightly lower error for trapezoid rule-based integration and backward difference approximation estimates. A simplified workflow of how p-ath773 is integrated into the ORKA framework is shown in Figure 4.2C and a more detailed explanation is included in Materials and Methods. In summary, the p-ath773 model includes the four tissue models and tissue-tissue interactions, whereas the ORKA to perform dFBA is the approach used to simulate the model form one time point to the next. The simulations of the p-ath773 model has been advanced through several growth stages using time points for changes in growth stage taken from experimental data (Boyes et al., 2001), see Figure 4.5. Figure 4.5 highlights the time points spread out through the seven growth stages modeled here including seed germination, seed germination to leaf development transition, leaf development, leaf development to flower production transition, flower production, flower production to silique ripening transition, and silique ripening. Figure 4.5 further provides sketches of the *in silico* and *in vivo* representations for each of these growth stages. In the seed germination stage, the uptake of fatty acids, sugars, and amino acids from seed storage (endosperm, see Seed Germination stage in Figure 4.5) has been modeled as a rate of usage which results in all stored fatty and amino acids being depleted by the end of the seed germination to leaf development transition (Clauss & Aarssen, 1994). This rate has been determined such that it is constant in mmol/h (see Data S1) yet needed conversion to the mmol/gDW·h units used throughout the p-ath773 model. Therefore, the rate at which the endosperm is utilized is scaled by the gDW of the leaf tissue (as the leaf tissue is modeled as interacting directly with the endosperm). This scaling advantageously results in a gradual decrease of the rate of nutrients uptaken from the endosperm stores (in mmol/gDW·h), as would happen in a seedling as the plant mass begins to far exceed the mass of the endosperm. A 12:12 hour light:dark diurnal rhythm has been chosen to match experimental conditions for the studies on starch and sucrose storage/uptake dependence (Shipley & Vu, 2002). Diurnal metabolism affects the model at all growth stages except for Seed Germination, when the cotyledons (embryonic leaves) are shaded from light by the soil and/or seed coat. In growth stages when plant tissue ratios are constant (i.e., the vegetative stages such as Seed Germination through Leaf Development), the tissue mass ratio values have been taken from values typical for herbaceous plants (0.511 gDW leaf:0.0.267 gDW root:0.211 gDW stem after adjusting from fresh weight to dry weight) (Clauss & Aarssen, 1994; Shipley & Vu, 2017; Baleja et al., 2015) (See Data S1). In growth stages when the ratios between tissues change (Boyes et al., 2001) (i.e., seed production or dispersion stage), a linear biomass "slider" is used, where a single parameter, seeding (s), is used to progress tissue mass ratios (see Figure 4.5). This ranges from $s = 0$ (normal vegetative tissue mass ratios) to $s = 1$ (mass ratios when maximum amount of seeds have been produced and have not yet been dispersed) and is linearly incremented from the point at which the first flower is produced to when all flowers are produced then decremented to when all silique (seed pods) are shattered, thus dispersing all seeds (see Data S1). A workflow showing how ORKA is applied to the p-ath773 model can be found in Figure 4.2C. In addition to using ORKA to perform dFBA, Flux Variability Analysis (FVA) (Johnson, Barbour, & Weyers, 2007) has been performed, at twelve points throughout the *Arabidopsis* lifecycle, selected to represent each growth stage and diurnal status in those stages (save the Leaf Development to Flower Production transitions which includes a single time point, see Figure 4.5), subject to all growth constraints, and a growth rate equivalent to the optimal growth rate to evaluate the variability in the balanced flux estimates. Flux Variability Analysis is performed at 1 Hour(s) After Germination (HAG, seed germination stage, dark), 70 HAG (seed germination to leaf development transition, light), 90 HAG (seed germination to leaf development transition, dark), 177 HAG (leaf development stage, light), 181 HAG (leaf development stage, light), 770 HAG (flower production stage, light), 810 HAG (flower production stage, dark), 1155 HAG (flower production to silique ripening transition, light), 1170 HAG (flower production to silique ripening transition, dark), 1190 HAG (silique ripening stage, dark), and 1199 HAG (silique ripening stage, light). In summary, we incorporated the p-ath773 model in an ORKA framework to simulate *Arabidopsis* metabolism across the lifecycle of an individual plant.

4.3.5. Design-build-test cycling of the p-ath773 model in the ORKA framework

Once growth stages have been implemented with the p-ath773 model and the ORKA framework, the design-build-test cycle (shown in Figure 4.3) has been used to iteratively improve and refine the p-ath773 model. The data points used to determine how well the model fits experimental literature include the mass of the whole plant at certain benchmark times and peak mass yields of leaf, seed, and stem tissues (Boyes et al., 2001; Shipley & Vu, 2002). At 17, 24, and 31 Days After Germination (DAG) the total dry plant mass should be between 0.5 and 2.0 mg; 2 and 8 mg; and 10 and 30 mg, respectively (Shipley & Vu, 2002). Upon the completion of multiple iteration of design-build-test cycle, the p-ath773 model has been adequately refined, the p-ath773 model has shown a total dry plant mass of 0.676 mg at 17 days (408 hours), 4.20 mg at 24 days (576 hours), and 25.9 mg at 31 days (744 hours) after germination. Furthermore, mass-based growth targets include the peak dry weights of the leaves, the seeds, and the stems which have been reported as approximately 163.7 mg (standard deviation 52.0 mg), 127.9 mg (standard deviation 52.7 mg), and 188 mg (standard deviation 39.3 mg), respectively (Boyes et al., 2001). As the path773 captures both plant growth and loss of seed (and other) mass in the silique ripening stage, the peak mass of each of these tissues has been comparted to this data. In the refined p-ath773 model, the peak masses of the leaves, seeds, and stems have been determined as 153 mg, 100 mg, and 151 mg, respectively, all of which are within one standard deviation of the experimental value (Boyes et al., 2001) (see the methods section for how tissue masses are determined). These comparisons are summarized in Figure 4.6. In summary, *in silico* tissue and plant mass values are similar to *in vivo* data, thus showing strong agreement with respect to plant- and tissue- scale growth trends. This agreement has been achieved by tuning the rate of carbon dioxide and light availability to the plant system (Shipley & Vu, 2002; Solovchenko & Merzlyak, 2008) which the modeled plant is allowed to utilize as well as by tuning the plant biomass yield (defined as the fraction of plant growth that adds to the plant mass with the remainder addressing litter, tissue repair, and degradation) (Thornley & Cannell, 1999; Cannell & Thornley, 1999). We have defined both carbon dioxide and light uptakes based on literature, with the former from the carbon assimilation rate (Li, Suzuki, & Hara, 1998) and the Leaf Area Ratio (LAR) of *Arabidopsis* (Sengupta & Majumder, 2014) and the latter from the transmission spectrum of fluorescent light bulbs (used in *in vivo* experiments utilized in the p-ath773 model reconstruction) (Baleja et al., 2015), the absorption spectra of chlorophyll (Baleja et al., 2015), and the Leaf Area Ratio of *Arabidopsis* (Li et al., 1998). However, the value of biomass yield (for a given plant across its full lifecycle) has been experimentally identified as between 0.7 and 0.85 (Thornley & Cannell, 1999). Here, to achieve the best alignment between *in silico* and *in vivo* growth patterns, biomass yield has been defined as 0.51. There are several possible reasons which are included in the Discussion section. All files necessary for p-ath773 have been included in the GitHub p-ath773 repository (DOI: 10.5281/zenodo.3735103). The *in silico* results of the final p-ath773 model can be found in Data S2.

4.3.6. Flow of water across plant lifecycle

Important to the life of a plant is the flow of water. Water carries various dissolved nutrients for transport (sugars, amino acids, nitrates, sulfates, et cetera) in addition to meeting the metabolic needs (such as photosynthesis) and physiological needs (such as transpiration) of tissues. Water flow through the plant has been selected as a case study which shows tissue-level insight into the general metabolic and transport processes modeled in p-ath773. The results of this analysis are shown in Figure 4.7, where each bar graph represents a specific stage of growth as shown in Figure 4.5. As can be seen in Figure 4.7, the stem tissue is the center of water transport, accepting water from the root and its own metabolism, and transporting this water to the leaf for its use in photosynthesis and to meet the physiological demands imposed upon the leaf by transpiration in addition to transportation to the seed tissue to meet its metabolic demands. Arrowheads indicate the most common direction of water flow, and negative reaction flux indicates flow in the opposite direction. The p-ath773 model shows that the primary driving force pulling water through the plant is transpiration, and that this driving force results in water flow rates during the light periods of two orders of magnitude higher than that which occurs in the dark periods. This *in silico* observation replicates the physiological water potential gradient along which water flows in plants which is driven by transpiration (Goldstein et al., 1998). Further, the pattern of water flow in the stem tissues being orders of magnitude higher during periods of light is consistent with *in vivo* data of other plant species[43]. While the role of transpiration in plant hydraulics is well known, the p-ath773 model framework in conjunction with the ORKA provides the opportunity to study the contribution of metabolic water to the flow of water in the plant system. In general, as modeled by p-ath773, it appears that root, stem, and seed tissues take up water and utilize it for their own metabolism, acting as water “sinks”. The leaf is however the largest water “sink” in the system since larger amount of water is transpired by the leaf tissue in comparison to that is used by the metabolism of other tissues. However, the leaf cytosol is a net producer of metabolic water, and the water transported from the cytosol to the extracellular compartment where transpiration is modeled to occur contributes between 60% and 80% of water which is transpired. Major metabolic contributions to the cytosolic water pool appear to be related to various metabolic processes not contained in other tissue models such as nitrate reduction, fatty acid metabolism, and a large number of other metabolic transactions which involve water.

4.3.7. Sulfur metabolism across plant lifecycle

In addition to tracking the flow of water through the plant, the p-ath773 model has also been used to study and track sulfur metabolism and transport across the tissues and the lifecycle of the plant to provide an example of reaction-level window into the p-ath773 modelled plant metabolism. This has been done to provide unique insight into the core metabolism of a single micronutrient which is not as extensively studied as carbon and nitrogen metabolism (Shaw & Cheung, 2018; Simons et al., 2014; L. Zhang et al., 2010), yet sulfur still is important to plant growth. The results of this analysis are shown in Figs. 4.8 and 4.9, where the former reports mean reaction rates and the latter reports mean concentrations for each specific stage of growth as shown in Figure 4.5. Sulfur is modeled as passing through the root and stem tissue and being distributed to the leaf and stem tissues. Some sulfur which has been distributed to the leaf tissue will be returned back to the stem, in the form of amino acids for distribution to the seed tissue, with the remainder being used to produce biomass. The seed accepts amino acids and sulfate from the stem tissue to produce biomass. Here it is evident that, in terms of sulfur metabolism, the seed serves as a "sink" tissue, the root as a "source" tissue, and the leaf as an intermediary. As is shown in Figure 4.8, the demand by the plant for sulfur is highest in the latter three stages of growth, where seed tissue is present and growing rapidly, or being loosed and metabolic demand from the seed corresponds to the increased maintenance and senescence of the seed and leaf tissues. The presence of seed tissue as a sulfur "sink" also leads to a high flux rate through many reactions in the sulfur metabolism in the leaf as well as transport of sulfur-containing amino acids through the stem tissue. These observations are largely as expected. Unexpected results are those related to the generally high rate of flux through portions of the sulfur metabolism in the leaf during the seed germination growth stage, and the corresponding low fluxes through these pathways in the seed germination to leaf development transition. From closer observations of metabolite concentrations and reactions rates as shown in Figures 4.8 and 4.9 (see Data S3), it appears that there are seemingly random switches between production, storage, and consumption of various metabolites such as L-homocysteine, methionine, and cysteine in the leaf in the early growth stages.

At some places in the Figures 4.8 and 4.9 metabolic maps, there appear some metabolites which have no initial concentration, yet a high mean concentration in the first stage of growth and mean fluxes away from that metabolite. This seems counter-intuitive. For one such metabolite cysteine in the leaf tissue in the first 60 hours after germination (the seed germination stage), the reaction converting hydrogen sulfide to cysteine has positive flux (average positive flux of 2.72E-3 mmol/gDW·h) for 13 of those hours, negative flux (average negative flux of -1.69E-3 mmol/gDW·h) for 22 hours, and no flux for 25. It appears that the no flux points in particular are positioned such that they occur when cysteine concentration is high, skewing the mean concentration upward. Notably, when the stores are used, a number of negative flux rates occur in a row. This skews the average reaction rate downward. It is also shown that cytosolic and extracellular cysteine have high concentrations. This is achieved by near constant interchange of cysteine position through a proton antiport. In the first 60 hours after germination (the seed germination stage) this antiport flows in the direction of the extracellular space 21 of those hours (average flux 0.001337 mmol/gDW·h), in the direction of the cytosol 38 of those hours (average flux -0.00098 mmol/gDW·h), and has no flux only at the first hour when there is as of yet no concentration of cysteine in the cytosol.

4.4. DISCUSSION

In the current work, a novel Optimization- and explicit Runge-Kutta- based Approach (ORKA) to dynamic Flux Balance Analysis (dFBA) has been developed. Inspired by the Static Optimization Approach (SOA) to perform dFBA, it seeks to achieve higher levels of model accuracy and solution stability. ORKA differs from the SOA in that it replaces first-order Taylor-series approximations for biomass and concentration steps with Runge-Kutta- and Trapezoid rule-based integration. This provides lower error floors, from $O(h^2)$ in the SOA to $O(h^4)$ in the ORKA, depending on the Runge-Kutta method used in the ORKA. ORKA has been developed to be general enough that several different Runge-Kutta methods could be applied to biomass step estimates (Figs. 2A and 2B) dependent on the error level desired or which could be achieved in the modelled system.

As a test system for ORKA, a multi-tissue core metabolism stoichiometric model of *Arabidopsis thaliana* has been reconstructed (Figure 4.3), which includes individual leaf, root, seed, and stem tissues models with unique metabolic roles (Figure .42). This model, named p-ath773, has defined intra-tissue interactions, interactions with the environment, and certain growth-based parameters defined based on growth stage in an effort to model *Arabidopsis* growth across its lifecycle by defining several growth stages (Figure 4.5). Once p-ath773 has been reconstructed, ORKA has then been applied (Figure 4.2C) using Heun's Third Order Rule. When the p-ath773 model using the ORKA (to perform dFBA) is compared to another *Arabidopsis* model utilizing the SOA (to perform dFBA) (Shaw & Cheung, 2018), the p-ath773 model in theory has at least a three order of magnitude lower error floor due to the smaller step sizes, increased accuracy of the dFBA approach used, and inclusion of two more tissue types. However, similar comparison with the most recent dFBA work on *Arabidopsis* lifecycle (Shaw & Cheung, 2018) is not entirely possible since these models are quite different in structure, goals, and results. For instance, the mass of the plant for the 6 to 36 days window of time is quite different between p-ath773 and the model produced by Shaw & Cheung (2018) (see Data S2 for details). In addition, comparing the rate of glutamine synthase in p-ath773 to that of Shaw & Cheung (2018)' we find marginal agreement between the two models. One of the primary differences between the models is the direction of the flow of amino acids in the models. While Shaw & Cheung (2018), show nitrate flow from the root to the leaf and then amino acid flow from the leaf to the root, the p-ath773 model synthesizes some amino acids in the roots and those amino acids being transported to the leaf tissue for consumption. Therefore, the direction of amino acid flow is reversed which is similar to what is reported in literature (Tegeder & Hammes, 2018; Santiago & Tegeder, 2016; J. Thornley & Cannell, 2000). Further, as the biomass equations are different between the two models, the p-ath773 model has a greater demand for amino acids and nitrogen atoms in its biomass composition than does Shaw & Cheung (2018). Therefore, by these models having different biomass, different flows of nitrogen, and different biomass composition and demands, it is very difficult to make a worthwhile comparison between the two models on the basis of accuracy as the structure is so different without strongly adapting one model or the other to be more similar to the other. Even though the p-ath773 model lacks a similar model in literature for the purposes of comparison, possibly because different literature sources and goals are used in model reconstructions, it is certain that when ORKA will be applied to a modeling framework comprising of all major tissues and can recapitulate and analyze real plant phenotypes. Further, these differences do not invalidate one model or the other, but rather might consider different metabolic states due to different growth conditions, thereby representing the flexibility of biological systems. In future, OKRA can be applied either by developing more tissue models (e.g., stem and seed) and adding to Shaw and Cheung's model or extending the p-ath773 model to capture the secondary metabolism, and either approach, carefully informed by literature, could greatly add to knowledge of *Arabidopsis* metabolism.

Using the ORKA to perform dFBA, p-ath773 is able to simulate seven stages of *Arabidopsis* growth (Figure 4.5) and showed agreement with literature on plant-scale growth (Figure 4.6) and on some reaction-level metabolic characteristics such as transpiration being a driving force of water flow through the plant system (Figure 4.7). One point on which there is lesser agreement between p-ath773 and *in vivo* plant-scale data is biomass yield, which is 51% in the p-ath773 model but for most species the value is between 70% and 85% *in vivo* (Grafahrend-Belau et al., 2013). This disparity is likely due to a few factors. The first is that the literature *in vivo* data generally accounts for factors such as harvesting and animal grazing (J. H. M. Thornley & Cannell, 1999; Cannell & Thornley, 1999), which is beyond the scope of the p-ath773 model, allowing for more growth. Further, the metabolic costs of root exudates (metabolites exported by the root to support the root microbial community) are not modeled. This is another potentially considerable drain on plant resources which is not modeled in the p-ath773 model.

The modeled flux rates have been used to study the flow of water through the plant system, and in particular to investigate the contributions of metabolic water to that transpired (Figure 4.7) and to investigate the whole-plant core metabolism of sulfur (Figures 4.8 and 4.9). In the former case study, the p-ath773 model has showed that metabolic water may contribute significantly to the amount of water transpired, somewhere between 60% and 80% of the total, and that transpiration drives a strong diurnal pattern of water flow. We hypothesize that the metabolic contribution to the amount of water transpired *in vivo* is unlikely to be as significant as shown by the p-ath773 model but is still likely to make some contribution. This is because not all water dynamics are accounted for in the p-ath773 model, including factors such as the amount of water necessary to keep new biomass turgid (since what is modeled is dry weight not wet weight) and the amount of water produced or consumed by the plant's extensive secondary metabolism. This shortcoming is common to all SMs rather than to the p-ath773 model in particular, as all such models only model dry weight.

For the sulfur metabolism case study, it has been shown that part of the patterns of sulfur metabolism are as expected such as increased use of and metabolic demand for sulfur when the seed tissue is present. However, some unexpected behavior has also been observed such as higher fluxes through sulfur reactions and comparatively larger concentrations of sulfur-containing metabolites at early growth stages. It is nearly impossible to pinpoint a single cause for the unexpected metabolic behavior of the sulfur metabolism in the early growth stages. This is due to the links between sulfur and energy metabolisms, in that many steps use some type energy molecule. Sulfur metabolism is also closely linked to the proton budget of the plant, in that many transports are proton-coupled. Through links to both the energy metabolism and proton budget, sulfur metabolism is strongly connected with the rest of plant metabolism. Hypothetically, this unexpected metabolic behavior might therefore be advantageous to the plant in energy metabolism and the control of the flow of protons. Particularly in the first two growth stages when the seedling's endosperm and cotyledons are not fully utilized and are therefore providing some amino acids (though notably not cysteine or methionine), fatty acids, and sugars. As modeled, these stores interact with the extracellular space of the leaf tissue, and often require facilitated transport (usually proton-coupling) into the cytosol for use or catabolism. It is therefore possible that these unexpected behaviors aid in the transport of nutrients from the endosperm, by having standing pools of metabolites which participate in proton-coupled transport to better regulate the cell's proton budget. This hypothesis is supported by the fact that these unexpected metabolic behaviors are reduced in magnitude as the amount of nutrients uptaken from the endosperm are reduced, and indeed the concentration of metabolites such as cysteine sharply decrease. These unexpected behaviors then appear to cease all together when the endosperm is fully utilized. While the metabolic network of p-ath773 is too convoluted to prove this theory, it does highlight the usefulness of stoichiometric modelling to identify interactions which may be too complex to deduce through non-systems approaches.

While there are a number of constraints applied to the model, such as biomass yield; maintenance and senescence costs; and enforcing mass ratios between tissues, these constraints apply mostly to plant-scale behaviors. These behavioral constraints generally fall into two categories: whole-plant and tissue-tissue interactions. The former generally ensure that the pattern of modeled plant and tissue growth fits that of *in vivo* data. The latter generally ensure that mass balance is maintained when metabolites are transported between tissues since each flux rate is in units of mmol/gDW tissue·h and each tissue is of a different mass. Hence, such conversions are necessary. Other constraints which fall in the category of tissue-tissue interactions ensure that nutrient flow is in a logical and well-known direction (e.g. micronutrients travel up from the roots). Few constraints, with the exception of the enforced diurnal patterns of carbon storage, apply on the reaction rate- or metabolite concentration- levels, leaving a large number of system degrees of freedom at the micro-scale. Therefore, by constraining the macro-scale behavior to what is known, the p-ath773 model can be used to determine what is, or may be, occurring in the plant system with respect to reaction rates or metabolite concentrations. From the allowed uncertainty at the micro-scale level, a study of this level allows investigation of what metabolic processes support and explain the known macro-scale behavior.

This work provides the basis for much future development and sophistication, both in broadening the range of approaches which can be taken to dFBA, and in the potential to use p-ath773 as a basis for modeling other plant systems. Applying ORKA to perform dFBA may provide the framework for other step-by-step dFBA approaches utilizing other ODE solving methods such as Taylor Series, Linear Multistep, or even adaptive step size methods depending on the needs of the modeled system. The current p-ath773 model could be further sophisticated by adding the secondary metabolism of the plant system, which constitutes a significant portion of metabolism in many plant systems. Further, several simplifications have been made regarding tissues, particularly related to seed tissue, at present. For instance, the model currently assumes when the plant is flowering, that flower biomass and metabolism is roughly equivalent to that of the seed. While this results in a simpler model, this model cannot be used to investigate certain metabolic hypothesis such as the cost to the plant resulting from flower pigmentation, pollen, and nectar production. Future work will include developing models for other plant tissues, such as flowers. In addition, as this is a core carbon metabolism model, it is likely quite similar to the core metabolism of other plant systems. Therefore, the p-ath773 model can serve as a basis for the development of lifecycle models for other plant systems, particularly annual eudicots which are of agricultural interest, such as rice (*Oryza sativa*), potatoes (*Solanum tuberosum*), tomatoes (*Solanum lycopersicum*), and soybeans (*Glycine max*).

4.5. FIGURES

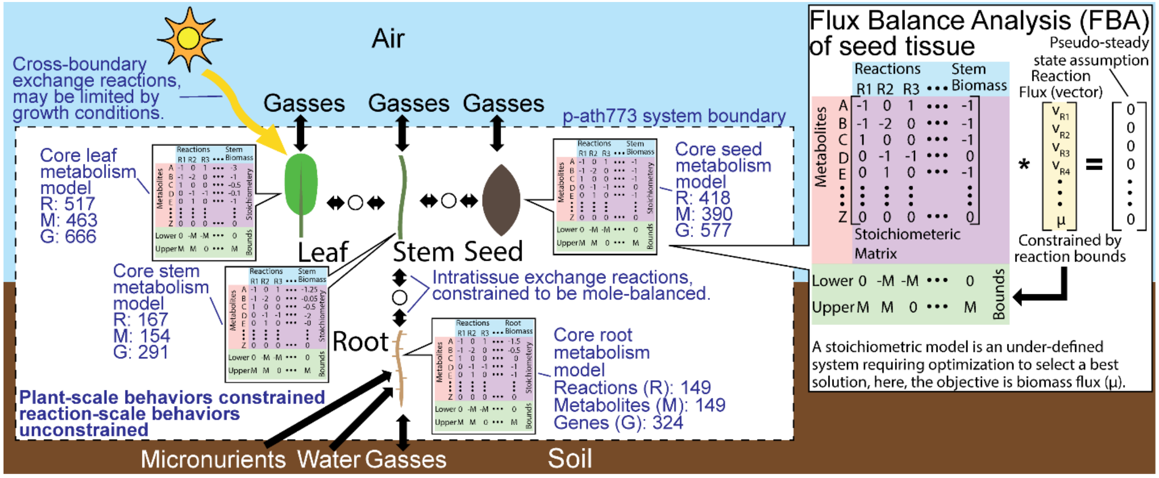
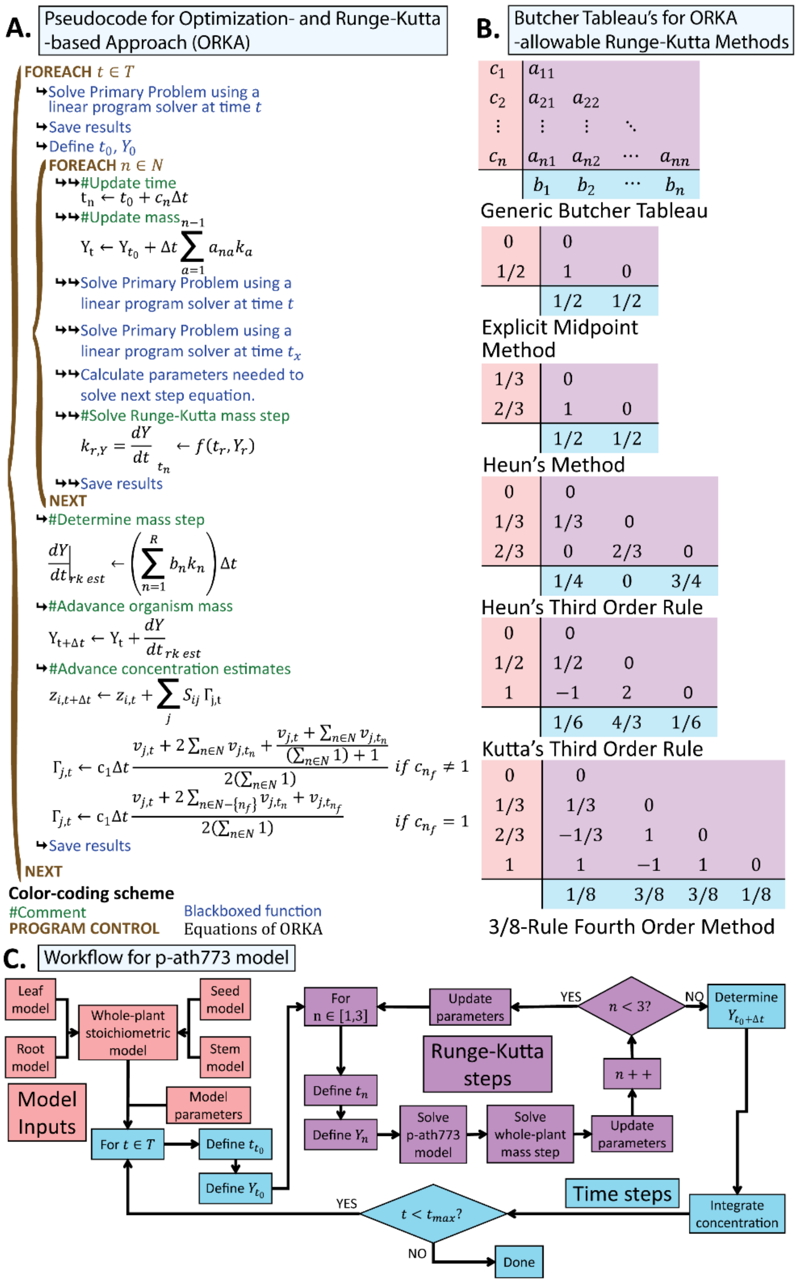

Air
Cross-boundary exchange reactions, may be limited by growth conditions.
Gasses
Gasses
Gasses
p-ath773 system boundary
Core leaf metabolism model
R: 517
M: 463
G: 666
Leaf
Stem
Seed
Core seed metabolism model
R: 418
M: 390
G: 577
Core stem metabolism model
R: 167
M: 154
G: 291
Intratissue exchange reactions, constrained to be mole-balanced.
Root
Core root metabolism model
Reactions (R): 149
Metabolites (M): 149
Genes (G): 324
Plant-scale behaviors constrained reaction-scale behaviors unconstrained
Micronurients
Water
Gasses
Soil
Reactions
Stem Biomass
Root Biomass
R1 R2 R3 •••
Metabolites
Stoichiometery
Lower
Upper
Bounds
Flux Balance Analysis (FBA) of seed tissue
Pseudo-steady state assumption
Reaction Flux (vector)
Stoichiometeric Matrix
Constrained by reaction bounds
A stoichiometric model is an under-defined system requiring optimization to select a best solution, here, the objective is biomass flux (μ).

Figure 4.1: The p-ath773 system model.

Extended Caption: This figure emphasizes the individual nature of each of the four core tissue models (leaf, root, seed, and stem), formally defines the modeled system boundary (dashed black line), defines cross-boundary exchange reactions, intra-tissue exchange reactions, and gives the generic formulation for Flux Balance Analysis applied to the seed tissue model.



A. Pseudocode for Optimization- and Runge-Kutta -based Approach (ORKA)

```
FOREACH $t \in T$
  ↪Solve Primary Problem using a linear program solver at time $t$
  ↪Save results
  ↪Define $t_0$, $Y_0$
  FOREACH $n \in N$
    ↪↪#Update time
      $\mathrm{t_n} \leftarrow t_0 + c_n \Delta t$
    ↪↪#Update mass
      $\mathrm{Y_t} \leftarrow \mathrm{Y}_{t_0} + \Delta t \sum_{a=1}^{n-1} a_{na} k_a$
    ↪↪Solve Primary Problem using a linear program solver at time $t$
    ↪↪Solve Primary Problem using a linear program solver at time $t_x$
    ↪↪Calculate parameters needed to solve next step equation.
    ↪↪#Solve Runge-Kutta mass step
      $k_{r,Y} = \frac{dY}{dt}\Big|_{t_n} \leftarrow f(t_r, Y_r)$
    ↪↪Save results
  NEXT
  ↪#Determine mass step
    $\frac{dY}{dt}\Big|_{rk\ est} \leftarrow \left(\sum_{n=1}^{R} b_n k_n\right) \Delta t$
  ↪#Adavance organism mass
    $\mathrm{Y_{t+\Delta t}} \leftarrow \mathrm{Y_t} + \frac{dY}{dt}_{rk\ est}$
  ↪#Advance concentration estimates
    $z_{i,t+\Delta t} \leftarrow z_{i,t} + \sum_j S_{ij} \Gamma_{\mathrm{j,t}}$
    $\Gamma_{j,t} \leftarrow \mathrm{c_1} \Delta t \frac{v_{j,t} + 2\sum_{n \in N} v_{j,t_n} + \frac{v_{j,t} + \sum_{n \in N} v_{j,t_n}}{(\sum_{n \in N} 1) + 1}}{2(\sum_{n \in N} 1)}$   $if\ c_{n_f} \neq 1$
    $\Gamma_{j,t} \leftarrow \mathrm{c_1} \Delta t \frac{v_{j,t} + 2\sum_{n \in N - \{n_f\}} v_{j,t_n} + v_{j,t_{n_f}}}{2(\sum_{n \in N} 1)}$   $if\ c_{n_f} = 1$
  ↪Save results
NEXT
```

Color-coding scheme

#Comment Blackboxed function
PROGRAM CONTROL Equations of ORKA

B. Butcher Tableau's for ORKA -allowable Runge-Kutta Methods

c_1	a_{11}			
c_2	a_{21}	a_{22}		
$\vdots$	$\vdots$	$\vdots$	$\ddots$	
c_n	a_{n1}	a_{n2}	$\cdots$	a_{nn}
	b_1	b_2	$\cdots$	b_n

Generic Butcher Tableau

0	0	
1/2	1	0
	1/2	1/2

Explicit Midpoint Method

1/3	0	
2/3	1	0
	1/2	1/2

Heun's Method

0	0		
1/3	1/3	0	
2/3	0	2/3	0
	1/4	0	3/4

Heun's Third Order Rule

0	0		
1/2	1/2	0	
1	−1	2	0
	1/6	4/3	1/6

Kutta's Third Order Rule

0	0			
1/3	1/3	0		
2/3	−1/3	1	0	
1	1	−1	1	0
	1/8	3/8	3/8	1/8

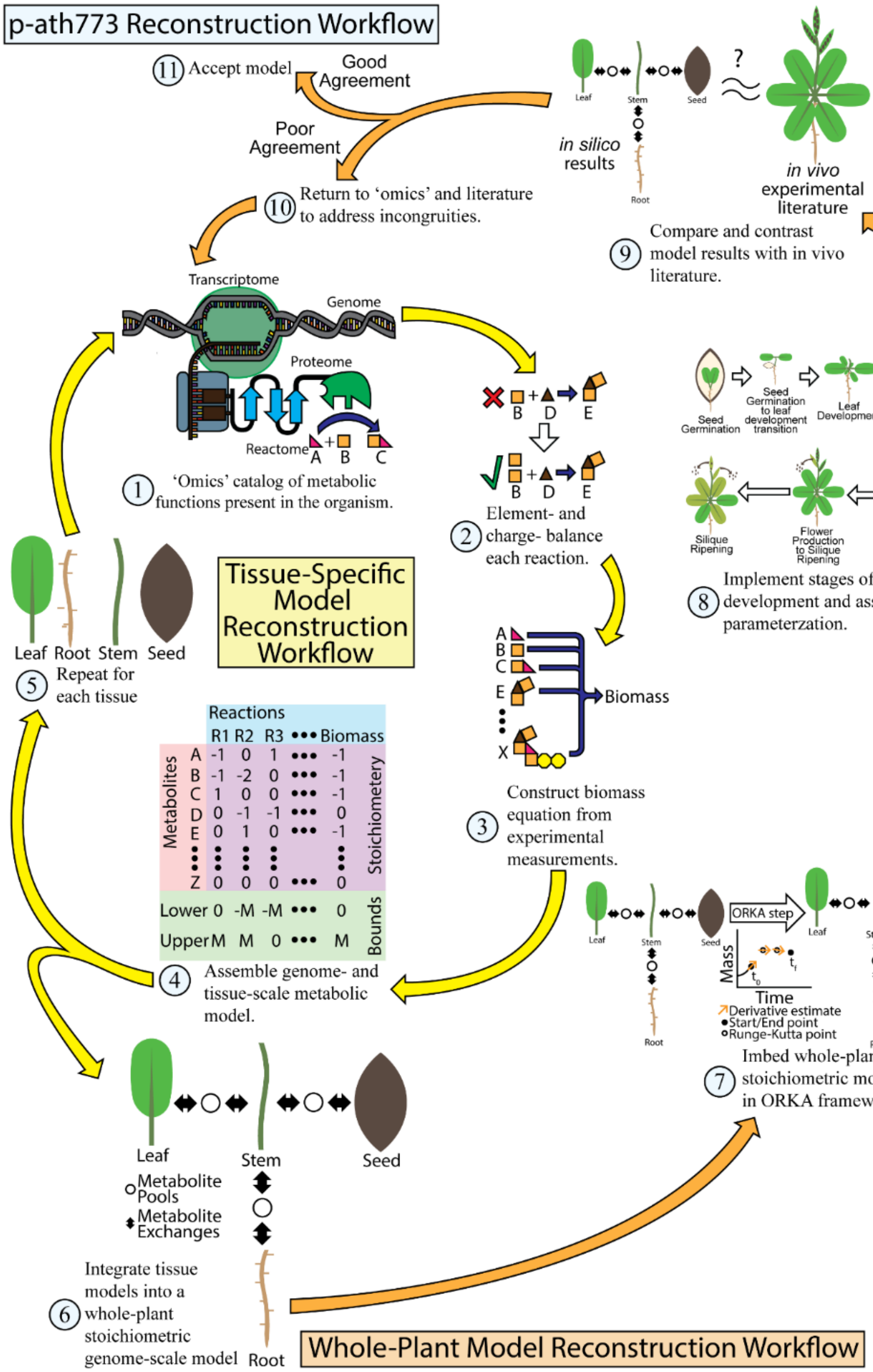
3/8-Rule Fourth Order Method

C. Workflow for p-ath773 model

Figure 4.2: Pseudocode, acceptable Runge-Kutta methods, and workflow with p-ath773 related to ORKA.

Extended Caption: This figure shows simple pseudocode appropriate to the implementation of the generic ORKA method in (A), Runge-Kutta method appropriate for use with the ORKA method in (B), and the workflow used in the specific application of ORKA to the p-ath773 model in (C). Symbols are defined as follows: t_n is the current time, t_0 is the initial time, Δt is the time step, c_n are the steps in the independent variable made by the Runge-Kutta method chosen to use, Y_t is the current biomass concentration, Y_0 is the initial biomass concentration, a_{na} is the weight of Runge-Kutta derivative estimate steps (k_n) in the next derivative estimate, b_n is the weight of the Runge-Kutta derivative estimate steps in the full Runge-Kutta derivative estimate, $\left.\frac{dY}{dt}\right|_{rk\ est}$ is the Runge-Kutta derivative estimate for the current timestep, $Y_{t+\Delta t}$ is the biomass concentration at the next time step, $z_{i,t}$ is the concentration of metabolite i at time t, $z_{i,t+\Delta t}$ is the concentration of metabolite i at the next time step, S_{ij} is the stoichiometric coefficient of metabolite i in reaction j, $\Gamma_{j,t}$ is the trapezoid rule-based integral estimate of the flux of reaction j at the current timestep, $v_{j,t}$ is the rate of reaction j at time t, v_{j,t_n} is the rate of reaction j at Runge-Kutta time step n, set N is the number of steps in the Runge-Kutta solution method (with n as the index), and c_{nf} is the final c_n value in the Runge-Kutta method. Greater detail on the definition of each symbols used can be found in the "Symbols Used" section in Text S1. In (A), there are two control loops (brown text with brown left-handed braces), one looping over each time point in the set of times over which to apply ORKA ($t \in T$), and the other looping over each step in the selected Runge-Kutta method ($n \in N$). The former control loop is used to solve the model at the time point, define the starting points for the Runge-Kutta method, and, after the Runge-Kutta loop is finished, advance biomass and metabolite concentrations in the model. The inner control loop determines the values of the Runge-Kutta-based concentration and biomass step estimates. The various estimates used rely on evenly spaced points at which the estimates are made, limiting the selection of Runge-Kutta method. Some allowable Runge-Kutta methods are shown in (B). For this work, Heun's Third Order

Rule was selected. In (C), an overview of the workflow used to integrate the p-ath733 model (red) in the ORKA method (blue and purple) is shown.



(11) Accept model

(10) Return to ‘omics’ and literature to address incongruities.

(9) Compare and contrast model results with in vivo literature.

(1) ‘Omics’ catalog of metabolic functions present in the organism.

(2) Element- and charge- balance each reaction.

(8) Implement stages of plant development and associated parameterzation.

(5) Repeat for each tissue

(3) Construct biomass equation from experimental measurements.

		Reactions				
		R1	R2	R3	•••	Biomass
Metabolites	A	-1	0	1	•••	-1
	B	-1	-2	0	•••	-1
	C	1	0	0	•••	-1
	D	0	-1	-1	•••	0
	E	0	1	0	•••	-1
	⋮	⋮	⋮	⋮		⋮
	Z	0	0	0	•••	0
Bounds	Lower	0	-M	-M	•••	0
	Upper	M	M	0	•••	M

Stoichiometery

(4) Assemble genome- and tissue-scale metabolic model.

(7) Imbed whole-plant stoichiometric model in ORKA framework

(6) Integrate tissue models into a whole-plant stoichiometric genome-scale model

Figure 4.3: Workflow for p-ath773 model reconstruction.

Extended Caption: This figure shows the workflow used in the reconstruction and curation of individual tissue models (yellow arrows) and the integrated p-ath773 model as a whole (orange arrows). The reconstruction procedure begins by consulting published ‘omics’ data which helps identify which metabolic functions are present in a given tissue, followed by element- and charge-balancing the reactions representing those functions. A biomass equation is defined from literature evidence, and a stoichiometric model of the reconstruction is created. This is repeated for each tissue until a plant-scale model can be created. This model is then placed in the ORKA framework, and is used to simulate plant growth throughout its lifecycle. The results are compared with *in vivo* experimental results, such as those shown in Figure 4.6. Incongruities are addressed at the tissue-level by re-consulting 'omics' level data. This process is repeated until an acceptable model is achieved.

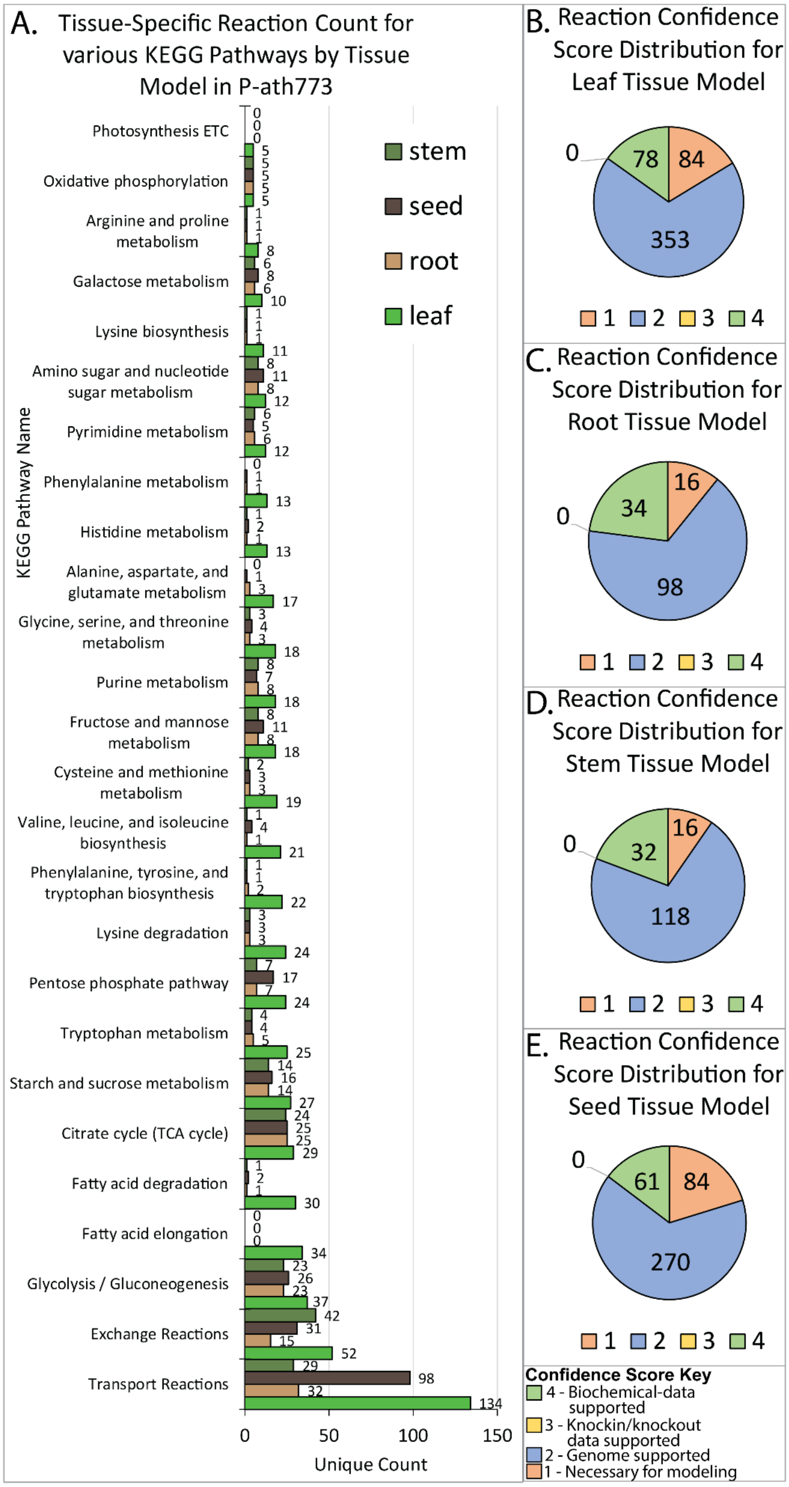


Figure 4.4: Statistics of tissue stoichiometric model reconstructions.

Extended Caption: Shown here are statistics related to the reconstruction of the leaf, root, seed, and stem models. (A) shows the types of reactions included in each of the four tissue models by counting the number of transport reactions, exchange reactions, and categorizing the remaining reactions based on the KEGG pathway(s) to which they belong. As shown here, the leaf model is the most complete and contains the most reactions is almost every category. Importantly, the leaf is the only tissue which contains reactions related to the photosynthetic electron transport chain (labeled "Photosynthesis ETC"). Figures (B) through (E) shows the rational for the inclusion of each reaction in each model using confidence scoring (see Thiele and Palsson for a definition and discussion of confidence scores). To summarize these figures, most reactions are included because there is evidence in the genome for these metabolic functions. The next most common reason for inclusion is being supported by biochemical literature data (e.g. a study has specifically identified the protein and determined its mechanism). The next most common reason for inclusion was modelling necessity (score of 1). No knock-in/knock-out studies where consulted in this work (score 3).

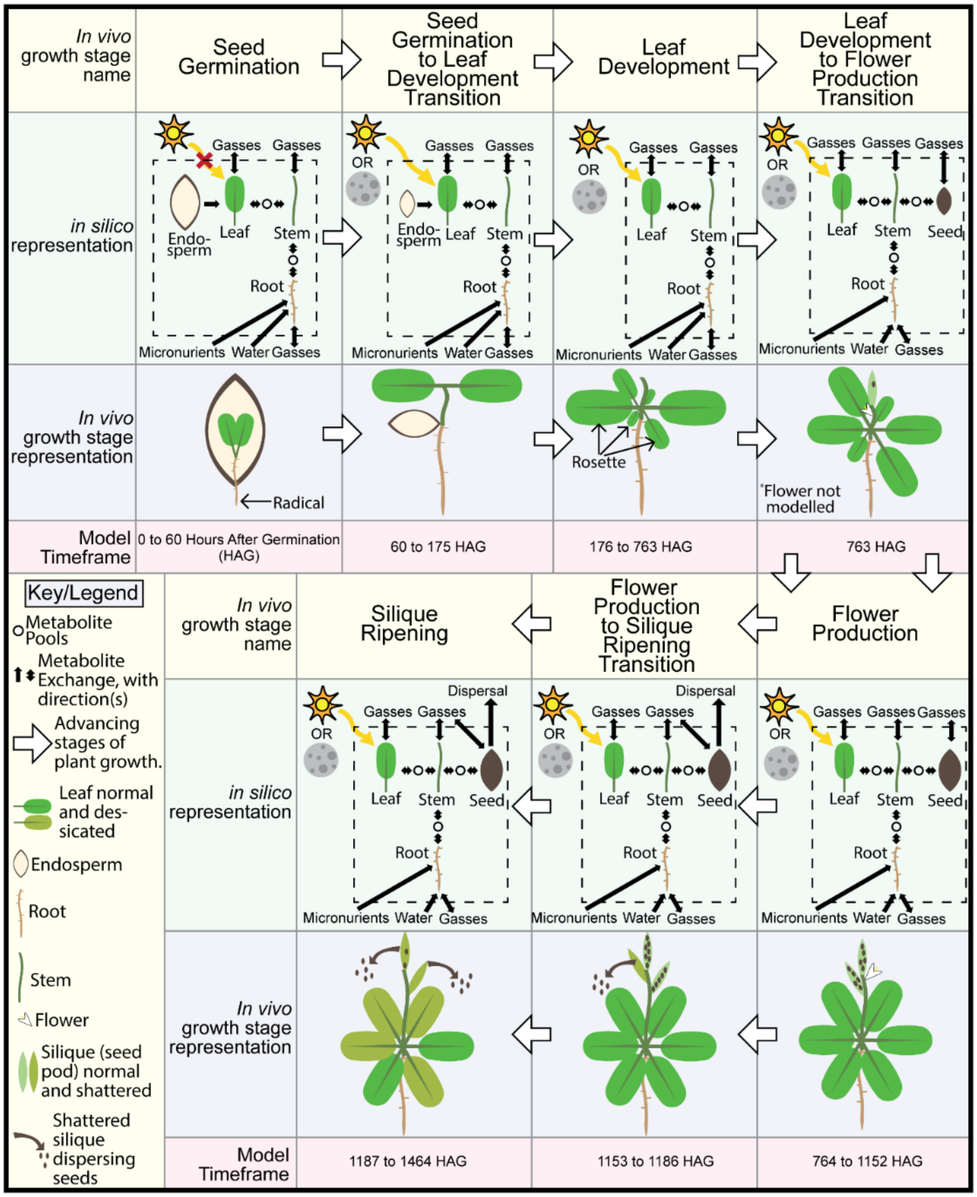

In vivo growth stage name
Seed Germination
Seed Germination to Leaf Development Transition
Leaf Development
Leaf Development to Flower Production Transition
in silico representation
Gasses
OR
Endo-sperm
Leaf
Stem
Seed
Root
Micronurients
Water
Dispersal
In vivo growth stage representation
Radical
Rosette
*Flower not modelled
Model Timeframe
0 to 60 Hours After Germination (HAG)
60 to 175 HAG
176 to 763 HAG
763 HAG
Flower Production
Flower Production to Silique Ripening Transition
Silique Ripening
764 to 1152 HAG
1153 to 1186 HAG
1187 to 1464 HAG
Key/Legend
Metabolite Pools
Metabolite Exchange, with direction(s)
Advancing stages of plant growth.
Leaf normal and des-sicated
Endosperm
Root
Stem
Flower
Silique (seed pod) normal and shattered
Shattered silique dispersing seeds

Figure 4.5: Seven growth stages in the p-ath773 model.

Extended Caption: Shown here are the labels given to each *in vivo* growth stage modeled *in silico* by p-ath773 (yellow headings), a sketch of the *in silico* representation (green rows) of the modeled plant system, a sketch of what the *in vivo* plant would look like at said growth stage (blue rows), and the timeframe in which the p-ath773 model simulates that growth stage as holding sway (red rows). White arrows indicate the progression of the system from germination to senescence. The *in silico* representation is a simplified drawing of what is occurring *in silico* showing major issue metabolite exchanges (black arrows), metabolite pools (open black circles) and interactions outside the system (black arrows crossing dashed-line box).

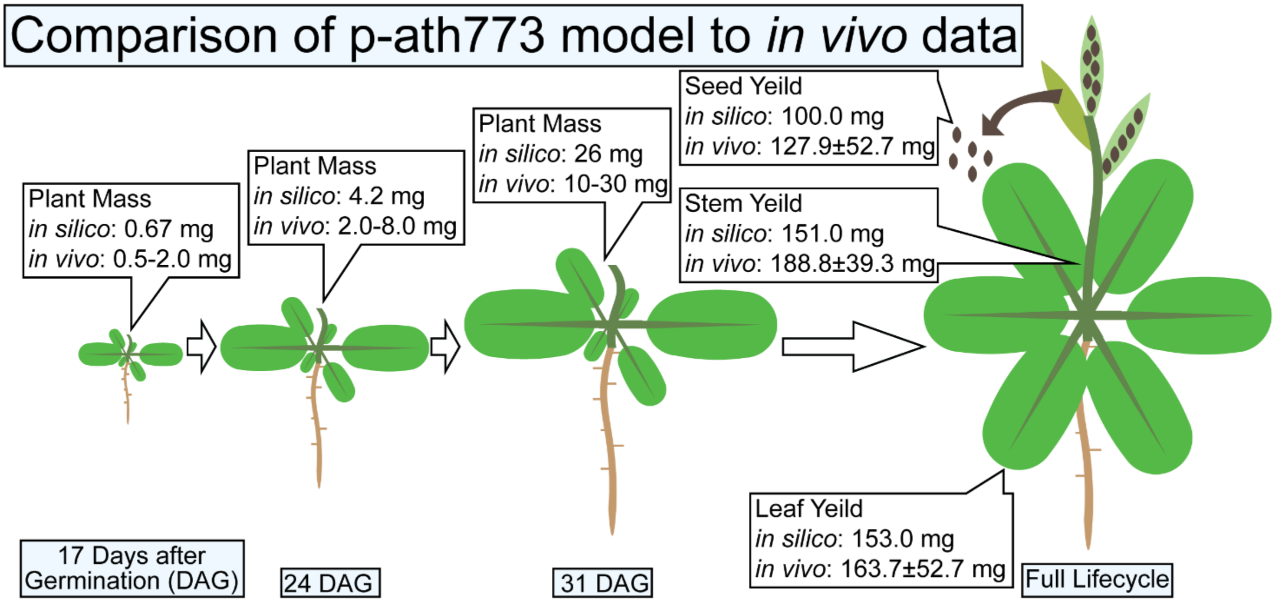


Figure 4.6: Comparison of plant-scale growth between *in vivo* data and the p-ath773 model.

Extended Caption: The figure shows some plant-scale growth check points which were used to verify the accuracy of the plant-scale growth pattern. The first three checkpoints were in the leaf development phase as 17 Days After Germination (DAG), 24 DAG, and 31 DAG, with *in vivo* experimental ranges for whole-plant mass and *in silico* whole-plant mass of the p-ath773 model shown in the callouts. The final image is for total tissue yield, where the reported *in silico* value is the maximum mass of each tissue during the entire lifecycle, and the *in vivo* value is the mean dry weight of the specified tissue at harvest plus or minus one standard deviation.

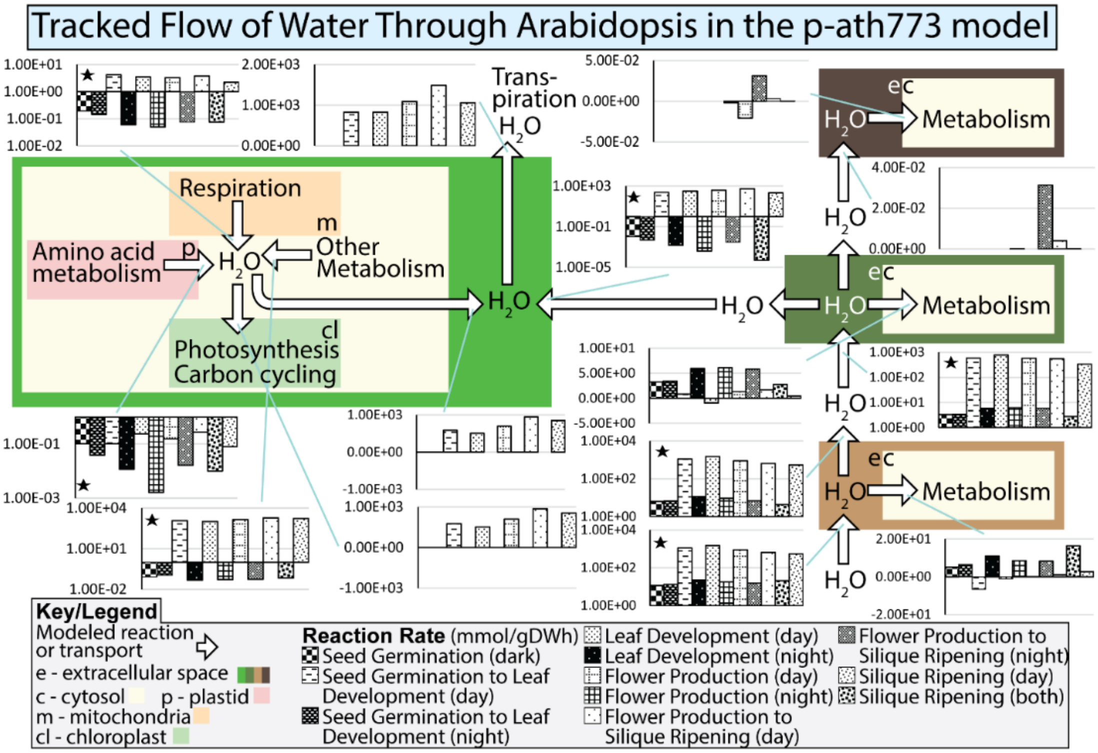


Figure 4.7: Tracked flow of water through *Arabidopsis* in the p-ath773 model.

Extended Caption: This figure shows the flow of water (white arrows) through the p-ath773 model by plotting the average reaction rate for each growth stage and each diurnal status of that growth stage, darker bars indicating growth at night and lighter bars indicating growth during the day, to highlight not only the stage-by-stage differences but also the diurnal differences. Flux rates are in units of mmol/gDW·h where gDW (grams dry weight) is in units of the dry weight of the individual tissue, rather than the plant as a whole causing incongruity as metabolites are exchanged between tissues as the flux rates must be scaled by the different tissues masses so none of a metabolite is gained or lost between tissue. Further, there are some hydrolysis reactions which occur in the extracellular compartment of each tissue, which accounts for the incongruity in the balance of water in tissue extracellular compartments (such as the in the seed tissue). This is generally a very small amount and therefore was not included in this figure. Further, logarithm-scale y-axes were used where possible (indicated by a small black star) because the day and night flux rates were generally orders of magnitude different.

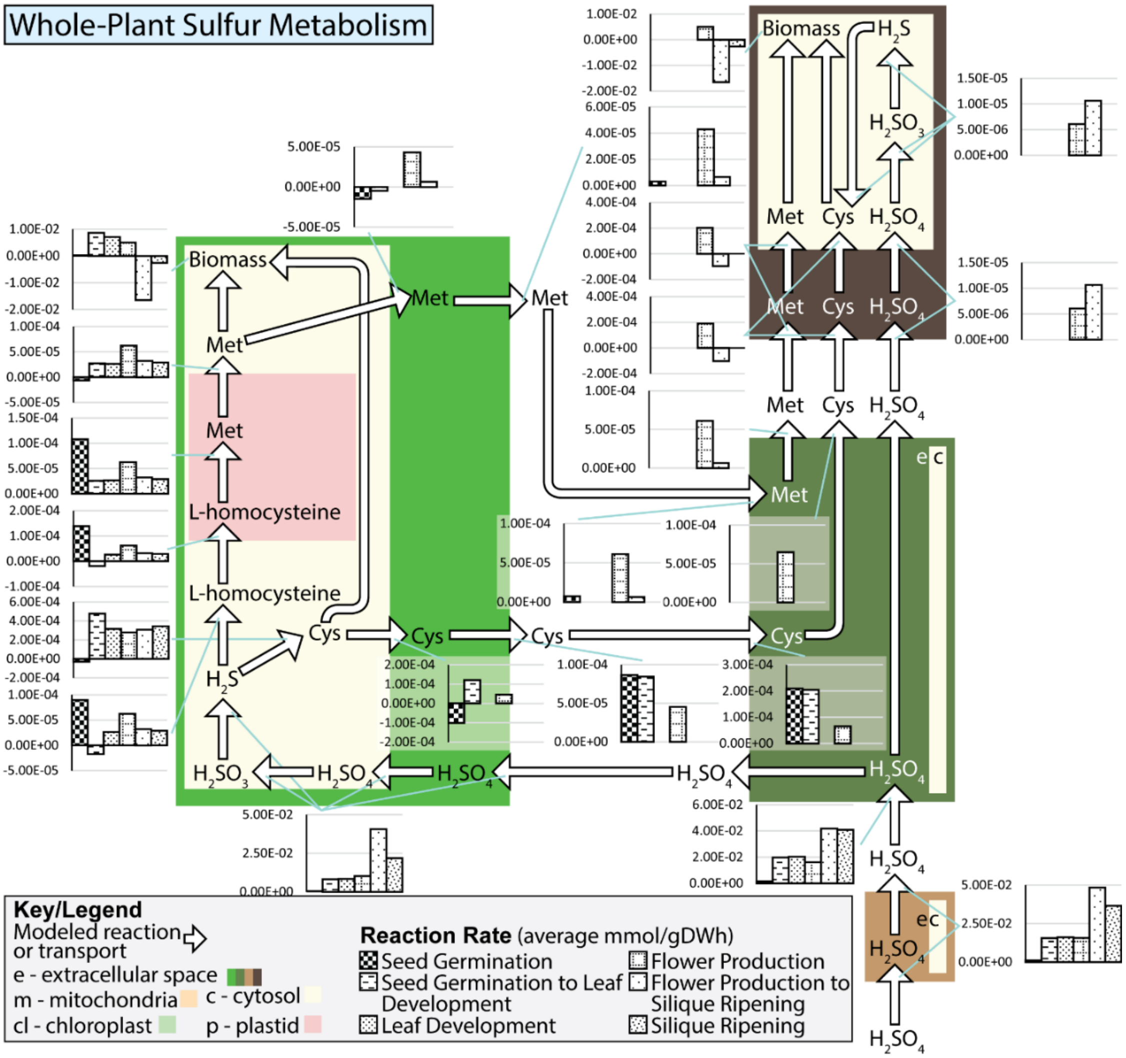


Figure 4.8: Rate of sulfur-utilizing reactions in Arabidopsis in the -ath773 model.

Extended Caption: This figure is meant to accompany Figure 4.9. This figure shows the evolution of the growth-stage mean reaction rates of reactions which transform or transport sulfur containing compounds in the p-ath773 model through the lifecycle of *Arabidopsis*. Flux rate values (black patterned bars) are in mmol/gDW·h, where gDW (grams Dry Weight) is in units of the dry weight of the individual tissue, rather than the plant as a whole causing incongruity as metabolites are exchanged between tissues as the flux rates must be scaled by the different tissues masses so none of a metabolite is gained or lost between tissue. Further, as sulfur-containing compounds are allowed to be stored in the model by building concentration, reaction rates may not balance.

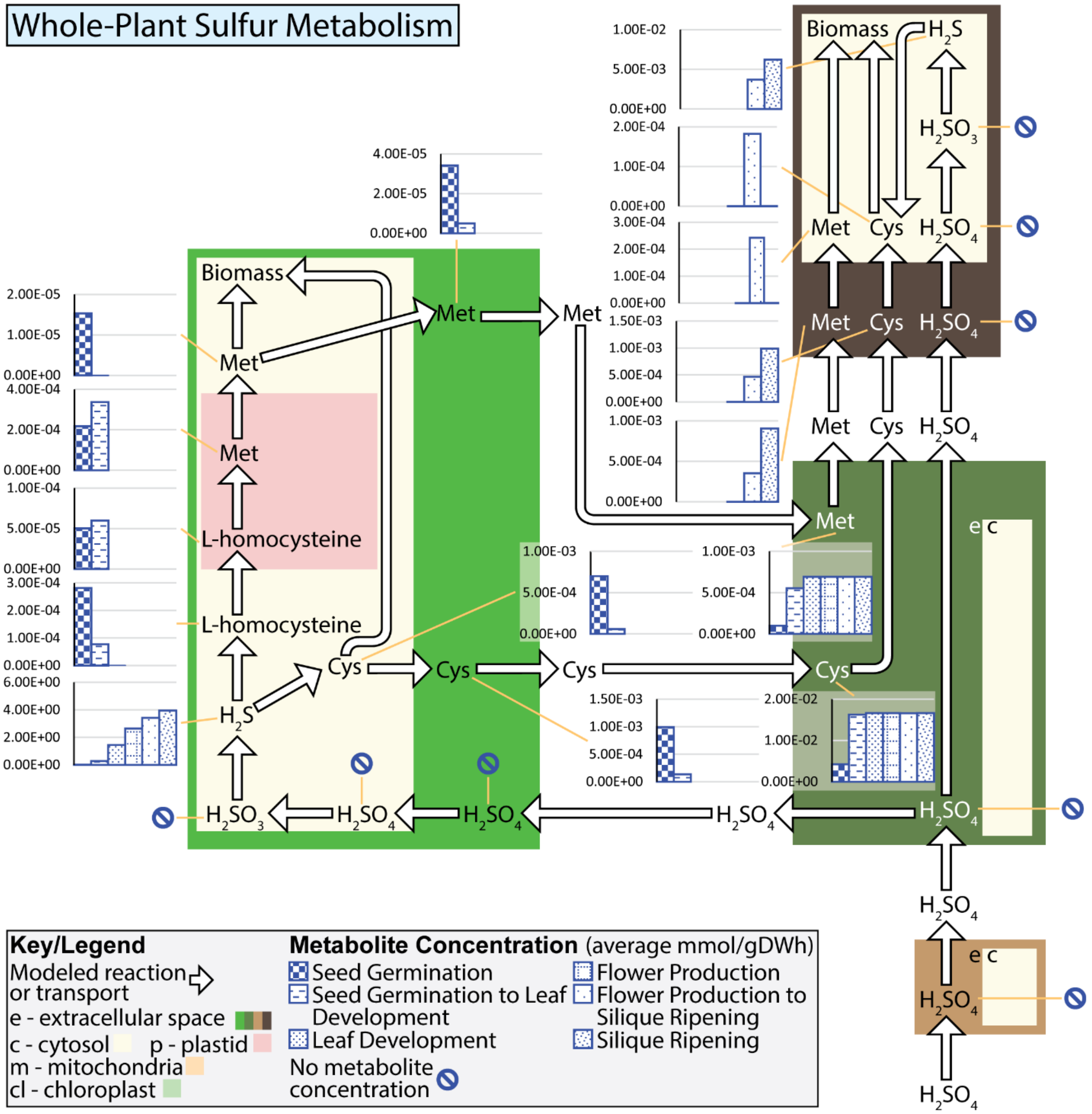


Figure 4.9: Concentration of sulfur-containing metabolites in Arabidopsis in the p-ath773 model.

Extended Caption: This figure is meant to accompany Figure 4.8. This figure shows the evolution of the growth-stage mean concentration of sulfur containing compounds in the p-ath773 model through the lifecycle of *Arabidopsis*. Concentration values (blue patterned bars) are in mmol/gDW.

4.6. MATERIALS AND METHODS

4.6.1. Development of the Optimization and explicit Runge-Kutta -based Approach to Perform dFBA

4.6.1.1. Static Optimization-Based dFBA Approach (SOA)

The Static Optimization-Based dFBA Approach (SOA) was first introduced in 2002 and is a method for solving for dynamic changes to a model system on a point-by-point basis (where those points are time), as opposed to the Dynamic Optimization-based Approach (DOA) which solves all points simultaneously (Mahadevan et al., 2002). The SOA, and variations thereon, have been applied to *Arabidopsis* (Shaw & Cheung, 2018) and barley (Grafahrend-Belau et al., 2009) plant models, making it of particular interest for the study of plant metabolism. The SOA is defined as follows (Mahadevan et al., 2002):

$$Maximize\ v_{biomass,t} \text{ (or some other suitable objective function)} \tag{4.1}$$

$$z_{i,t+\Delta t} \geq 0 \qquad \forall t \in T; i \in I' \tag{4.2}$$

$$v_{j,t}^{LB} \leq v_{j,t} \leq v_{j,t}^{UB} \tag{4.3}$$

$$z_{i,t+\Delta t} = z_{i,t} + \sum_{j} S_{ij} v_{j,t} \Delta t \qquad \forall t \in T; i \in I' \tag{4.4}$$

$$Y_{t+\Delta t} = Y_t + v_{biomass} Y_t \Delta t \qquad \forall t \in T \tag{4.5}$$

$$other\ constraints$$

Where symbols used are defined in the caption of Figure 4.2, in the Results section, and in the “Symbols Used” section of Text S1. Note that $v_{biomass}$ indicates the rate of reaction for the biomass production and is equivalent to the growth rate, μ_t which will be used hereafter. It should also be noted that biomass concentration, Y_t, is actually an element of the I' set (the set of metabolites whose concentration is tracked). However, equation (4.5) is included here for Y_t to be consistent with definitions of SOA in previous works (Mahadevan et al., 2002). This is also necessary due to the fact that biomass concentration is of particular interest in stoichiometric modeling efforts. Therefore, even though equation (4.4) simplifies to equation (4.5), equation (4.5) is still explicitly stated. This simplification is accomplished by first recognizing that there is only a single $S_{biomass,j}$ that is non-zero, namely $S_{biomass,biomass}$ which has a value of 1. Making the substitution, the RHS of equation (4.4) reduces to $z_{biomass,t} + v_{biomass} z_{biomass,t} \Delta t$. Secondly, by recognizing that Y_t is equivalent to $z_{biomass,t}$ and making this substitution on both sides of equation (4.4), equation (4.4) reduces to equation (4.5).

The mass step taken at each time point in the SOA method, as shown in equation (4.5), is derived from the Taylor series expansion of $e^{\mu x}$ around 0. The exponential formulation comes from the fact that the growth rate determined by a SM of metabolism is an exponential growth rate defined by the following differential equation[8].

$$\frac{dY_t}{dt} = \mu_t Y_t \tag{4.6}$$

Whose solution can be represented as follows:

$$Y_{t+\Delta t} = Y_t e^{\mu t} \tag{4.7}$$

To derive equation (4.5) that is used to advance biomass in the SOA, the first-order Taylor series expansion of $e^{\mu t}$ is used. Recall that the Taylor series expansion of $f(t)$ around point a is defined as follows.

$$f(t) = f(a) + f'(a)(t-a) + \frac{f''(a)}{2!}(t-a)^2 + \frac{f^{(3)}(a)}{3!}(t-a)^3 + \cdots + \frac{f^{(n)}(a)}{n!}(t-a)^n \tag{4.8}$$

It is well known that $f(x) = e^{\mu t}$ then $f^{(n)}(x) = \mu^n e^{\mu t}\ \forall n$, this Taylor series expansion may be simplified as follows.

$$f(t) = e^a + \mu e^a(t-a) + \frac{\mu^2 e^a}{2!}(t-a)^2 + \frac{\mu^3 e^a}{3!}(t-a)^3 + \cdots + + \frac{\mu^n e^a}{n!}(t-a)^n \tag{4.9}$$

Finally, since this Taylor series expansion is around $a = 0$, the taylor series expansion becomes what follows (knowing that $e^0 = 1$).

$$f(t) = 1 + \mu t + \frac{\mu^2 t^2}{2!} + \frac{\mu^3 t^3}{3!} + \cdots + + \frac{\mu^n t^n}{n!} \tag{4.10}$$

In the standard SOA formulation, only the first term is used in the approximation of $e^{\mu t}$, therefore:

$$g(t) = 1 + \mu t + O(t^2) \tag{4.11}$$

Multiplying through by Y_t gives the biomass steps which are used by the SOA, namely equation (4.5). Therefore, the error in mass step estimates in the standard SOA method is on the order of the timescale squared, e.g. $O(h^2)$. One additional assumption is made in equation (4.4).

The differential equation which describes the change in concentration of a specific metabolite is shown below.

$$\frac{dz_i}{dt} = \sum_{j \in J} S_{ij} v_{j,t} \tag{4.12}$$

By using the separation of variables as the solution method,

$$dz_i = \sum_{j \in J} S_{ij} v_{j,t} \, dt \tag{4.13}$$

$$dz_i = \sum_{j \in J} S_{ij} \int_{t_o}^{t_0+\Delta t} v_{j,t} dt \tag{4.14}$$

To derive equation (4.4) (used in the SOA) from equation (4.14), it must be assumed that for each time step Δt $v_{j,t}$ is constant.

4.6.1.2. Optimization- and explicit Runge-Kutta-based dFBA Approach (ORKA)

The dFBA method developed in this work is similar to SOA in the sense that both solve time points in a step-by-step and cumulative fashion. The Optimization- and explicit Runge-Kutta-based dFBA Approach (ORKA) differs in that different approximations are used in the solutions attempts to increase the accuracy of the estimation of the concentration and mass steps.

$$Maximize \ \mu_t \tag{4.1}$$

$$z_{i,t+\Delta t} \geq 0 \qquad \forall t \in [t_o, t_f]; i \in I' \tag{4.2}$$

$$v_{j,t}^{LB} \leq v_{j,t} \leq v_{j,t}^{UB} \tag{4.3}$$

$$z_{i,t+\Delta t} = z_{i,t} + \sum_{j} S_{ij}\,\Gamma_{\mathrm{j,t}} \qquad \forall t \in [t_o, t_f];\, i \in I' \tag{4.15}$$

$$Y_{t+\Delta t} = Y_t + \Delta t \frac{dY_t}{dt}_{rk\ est} \qquad \forall t \in [t_o, t_f] \tag{4.16}$$

$other\ constraints$

Where $\frac{dY_t}{dt}_{rk\ est}$ represents the Runge-Kutta based estimate for the mass step of the model, and $\Gamma_{\mathrm{j,t}}$ represents an estimate of the integral of $v_{j,t}$. Equation (4.15) expands on the accuracy of metabolic concentration estimates by leaving the integral term present and by removing the assumption that the reaction rate is time-independent in the time step concerned. This equation will estimate the integral using the multiple-applications Trapezoidal Rule of integration with the generalized explicit Runge-Kutta method. Equation (4.16) makes a more accurate estimate of the change in mass by using a Runge-Kutta method to better estimate the size of the mass step. With the notation altered to be more consistent with this work, the generic n^th^ order Runge-Kutta method is presented below.

$$\frac{dY}{dt} = f(t, Y_t) \tag{4.17}$$

$$Y_{t+\Delta t} = Y_t + \left(\sum_{n=1}^{N} b_n k_n\right)\Delta t \tag{4.18}$$

$$Y_{t+\Delta t} - Y_t = \left(\sum_{n=1}^{N} b_n k_n\right)\Delta t = \frac{dY}{dt}_{rk\ est} \tag{4.19}$$

$$k_1 = f(t, Y_t) \tag{4.20}$$

$$k_2 = f(t + c_2\Delta t, Y_t + a_{21}\mathrm{k}_1\Delta t) \tag{4.21}$$

$$k_3 = f(t + c_3\Delta t, Y_t + (a_{21}\mathrm{k}_1 + a_{32}k_2)\Delta t) \tag{4.22}$$

$$\vdots$$

$$k_n = f\left(t + c_n\Delta t, Y_t + \Delta t \sum_{j=1}^{i-1} a_{ij}k_j\right) \tag{4.23}$$

Where, for an explicit Runge-Kutta method, the above method constants (a_{nm}, b_n, and c_n) are represented in a triangular Butcher tableau such as shown below in Figure 4.2B. As previously stated, biomass growth is exponential as shown in equation (4.6).

$$\frac{dY_t}{dt} = \mu_t Y_t \tag{4.6}$$

Therefore:

$$\frac{dY_t}{dt} = f(t, Y_t) = \mu_t Y_t \tag{4.24}$$

However, since the value of μ_t is calculated by performing FBA on an SM, the SM must be solved for each k_n (Runge-Kutta derivative estimate for step n). Therefore, by this necessity, we would also know reaction rates at each time point where a k_n is solved for, namely, $v_{j,t}$, $v_{j,t+c_2\Delta t}$, ..., $v_{j,t+c_{n_f}\Delta t}$. Returning to equation (4.15), should the time points at which k_n occurs be equally spaced, then the multiple-application Trapezoidal Rule (i.e., the area of multiple trapezoids is used to estimate the integral rather than a single trapezoid) might be applied to solve the integral presented in equation (4.15). Evenly spaced points at which k_n values are calculated is defined that if $c_{n_f} - c_{n_f-1} = c_{n_f-1} - c_{n_f-2} = \cdots = c_2 - c_1$ is true, the points at which k_n are calculated are evenly spaced. Further, if $c_{n_f} \neq 1$, then $1 - c_{n_f} = c_2 - c_1$ must also be true. This may seem like highly specific criterion; however, several specific Runge-Kutta methods or rules fit these descriptions. These include the explicit midpoint method, Heun's Method, Heun's Third Order

Rule, Kutta's Third Order Rule, and the 3/8-Rule Fourth Order Method, among others. The Butcher tableaus for these methods can be found in Figure 4.2B. Using any of the above-mentioned methods, equation (4.15) could be restated as follows.

$$z_{i,t+\Delta t} = z_{i,t} + \sum_{j} S_{ij}\,\Gamma_{\mathrm{j,t}} \qquad \forall t \in T; i \in I' \qquad (4.25)$$

$$\Gamma_{j,t} = \mathrm{c}_1 \Delta t \frac{v_{j,t} + \sum_{n\in N-(n_f)}^{n-1} v_{j,t_n} + v_{j,t+\Delta t}}{2(\sum_{n\in N} 1)} \qquad (4.26)$$

$$t_n = t_0 + c_n \Delta t \qquad (4.27)$$

$$Y_{t_n} = Y_{t_0} + \Delta t \sum_{a=1}^{n-1} a_{na} k_a \qquad (4.28)$$

The only unknown quantity above is the Runge-Kutta estimate for the mass step ($v_{j,t+\Delta t}$). As the exact value of $z_{i,t+\Delta t}$ may be necessary to calculate $v_{j,t+\Delta t}$, it will be assumed that $v_{j,t+\Delta t}$ is equal to the arithmetic mean of the reaction rates in the n time points used in the Runge-Kutta method selected plus the starting point. Therefore, equation (4.26) may be rewritten as follows.

$$\Gamma_{j,t} = \mathrm{c}_1 \Delta t \frac{v_{j,t} + 2\sum_{n\in N} v_{j,t_n} + \frac{v_{j,t} + \sum_{n\in N} v_{j,t_n}}{(\sum_{n\in N} 1) + 1}}{2(\sum_{n\in N} 1)} \qquad if\ c_{n_f} \neq 1 \qquad (4.29)$$

Note that this correction to $\Gamma_{j,t}$ only applies when $c_{n_f} \neq 1$. Otherwise, the case when $c_{n_f} = 1$ removes the need for an arithmetic estimate of the final data point and the following equation might be used.

$$\Gamma_{j,t} = \mathrm{c}_1 \Delta t \frac{v_{j,t} + 2\sum_{n\in N-\{n_f\}} v_{j,t_n} + v_{j,t_{n_f}}}{2(\sum_{n\in N} 1)} \qquad if\ c_{n_f} = 1 \qquad (4.30)$$

The advantage of using a multiple application Trapezoidal rule in estimating the integral in equation (4.15) is that the assumption of a constant reaction rate over the time step may be relaxed and that the error for the multiple application trapezoidal rule is $O(\Delta t^3)$. From this, the ORKA method can be represented as follows.

For each time $t \in [t_0, t_f]$:

$$Maximize\ \mu_t \qquad (4.1)$$

$$z_{i,t+\Delta t} \geq 0 \qquad \forall i \in I' \qquad (4.2)$$

$$v_{j,t}^{LB} \leq v_{j,t} \leq v_{j,t}^{UB} \qquad \forall j \in J \qquad (4.3)$$

$$\frac{dY_t}{dt}_{rk\ est} = \left(\sum_{n=1}^{R} b_n k_n\right) \Delta t \qquad (4.19)$$

$$Y_{t+\Delta t} = Y_t + \Delta t \frac{dY_t}{dt}_{rk\ est} \qquad (4.16)$$

$$z_{i,t+\Delta t} = z_{i,t} + \sum_j S_{ij}\, \Gamma_{\mathrm{j,t}} \qquad \forall i \in I' \qquad (4.26)$$

$$\Gamma_{j,t} = (c_2 - \mathrm{c}_1)\Delta t \frac{2\sum_{n\in N} v_{j,t_n} - v_{j,t_{n_0}} + \frac{\sum_{n\in N} v_{j,t_n}}{\sum_{n\in N} 1}}{2(\sum_{n\in N} 1)} \qquad if\ c_{n_f} \neq 1 \qquad (4.29)$$

$$\Gamma_{j,t} = (c_2 - \mathrm{c}_1)\Delta t \frac{2\sum_{n\in N} v_{j,t_n} - v_{j,t_{n_0}} - v_{j,t_{n_f}}}{2(\sum_{n\in N} 1)} \qquad if\ c_{n_f} = 1 \qquad (4.30)$$

$other\ constraints\ such\ as\ mass\ balance\ and\ nutrient\ uptake\ which\ are$

$necessary\ for\ modeling$

For each Runge-Kutta Step $n \in N$, denoting the current time as t_c

$$k_n = f\left(t_n, Y_{t_n}\right) \qquad \forall n \in N \qquad (4.20)$$

$$t_n = t_c + c_n \Delta t \qquad \forall n \in N \qquad (4.27)$$

$$Y_{t_n} = Y_{t_0} + \Delta t \sum_{a=1}^{n-1} a_{na} k_a \qquad \forall n \in N \qquad (4.28)$$

To this point, the ORKA framework is deliberately general enough that any Runge-Kutta method which satisfies the criteria related to the values of c_1 through c_{n_f} might be selected. It should be noted that the multiple application Trapezoidal rule has an error floor in the order of $O\left((c_2 - c_1)h^3\right)$. Therefore, the integral estimate will have lower error than third-order Runge Kutta methods, which generically have a global error of $O(h^3)$ because of their use of individual steps of the Runge Kutta method rather than the full time step. The integral estimate will be the limiting accuracy factor if fourth order Runge-Kutta methods (such as the 3/8-rule fourth order method) or better are used which have a global error of $O(h^4)$.

4.6.1.3. Limitation of ORKA to explicit Runge-Kutta Methods

Generally, Runge-Kutta methods are defined as either implicit or explicit. A method is defined as explicit if $a_{ij} = 0 \; \forall \; j \geq i$, and implicit if this is not the case. Recall that a_{ij} defines the dependence of one derivative estimate step on another, as shown below.

$$k_n = f\left(t + c_n \Delta t, Y_t + \Delta t \sum_{j=1}^{i-1} a_{ij} k_j\right) \qquad (4.23)$$

If a method is explicit, each estimate depends only on previous estimates, whereas if a method is implicit, it may rely on future estimates or even itself. Implicit Runge-Kutta methods therefore are more difficult to implement, requiring the solution of a non-linear system of equations. In this work, $f(t, Y)$ is the solution of a large system of under-defined linear equations (the stoichiometric model) where the best solution is selected by optimization. Using an implicit Runge-Kutta method would require the full model to be included in an under-defined non-linear system of equations and then to be solved by non-linear programming approaches. These approaches neither guarantee that a solution would be found nor a solution found would be optimal (Kanehisa et al., 2017). The sharply increased computational costs of using an implicit method combined with the complexity of implementation and the non-guarantee of an optimal solution has made implicit Runge-Kutta methods not worthwhile or attractive for implementation.

4.6.2. Overview of the reconstruction of core metabolic models of leaf, root, seed, and stem tissues

The seed tissue has been modeled primarily based on a published MFA work (Lonien & Schwender, 2009) allowing an accurate reconstruction of the central carbon metabolism of the seed. Next, the leaf tissue has been reconstructed as a phototrophic tissue to supply carbon to the seed tissue. We next have reconstructed the root model to provide a mechanism for the uptake of water and micronutrients necessary for plant growth. Finally, we have reconstructed the stem model to provide a logical link between the tissues. Additional works that have been used in the reconstruction of tissue models can be found in Data S1.

4.6.2.1. The seed tissue model

The general workflow which has been used for the development of the four core tissue models is illustrated in Figure 4.3. The seed model has been developed first, with the central metabolic pathways based on a Metabolic Flux Analysis (MFA) of four seed genotypes published previously[29]. We then manually have filled gaps in this model with reactions based on literature and genomic evidence (Thiele & Palsson, 2010; Zomorrodi & Maranas, 2012) or with reactions being necessary for ensuring model connectivity. The stoichiometric coefficients of biomass precursors have been determined using sink reactions, dry biomass weight composition, and amino acid mass ratios provided in a previous work (Lonien & Schwender, 2009) (see Text S1). The resultant seed tissue model focuses on storage, respiration, and growth, and consists of 418 reactions, 577 genes, and 390 metabolites GitHub p-ath773 repository (DOI: 10.5281/zenodo.3735103) for this work.

4.6.2.2. The leaf tissue model

Next, we have reconstructed the leaf model by taking common reactions/pathways from the seed model and adding metabolic pathways for amino acids that are not synthesized in the seed. In addition, other leaf-specific pathways such as photosynthesis, carbon fixation, and gluconeogenesis and necessary transport reactions have also been added. We then have developed the biomass equation for the leaf tissue using that of a previously published *Arabidopsis* model (Saha et al., 2011) (see Text S1), with minor adjustments. First, since the p-ath773 model is designed to focus on core metabolism, secondary metabolites were removed from biomass equation. Second, it was noticed that the amino acid histidine was missing as a primary metabolite from biomass composition. Histidine was added into the leaf biomass in proportion to other amino acids (See Data S1) (Lonien & Schwender, 2009). The resultant leaf tissue model has focused on photosynthesis, respiration, gas exchange, fatty acid synthesis, and growth, and contains of 517 reactions, 666 genes, and 463 metabolites. We have included the leaf model in the GitHub p-ath773 repository (DOI: 10.5281/zenodo.3735103).

4.6.2.3. The root and stem tissue models

We have constructed the root and stem models, similarly, by extracting common reactions/pathways from the seed model and adding necessary and root-/stem-specific transport and exchange reactions. Then exchange reactions have been added to allow the root to be linked to micronutrient uptake processes from the soil and the stem to be involved in inter-tissue transport processes. In the absence of *Arabidopsis*-specific estimates, the dry weight composition of switchgrass (*Panicum virgatum*) root and stem (Johnson et al., 2007) have been assumed to be equivalent to the biomass composition of these tissues in *Arabidopsis*. Thus, we have found the biomass of root and stem tissues to be composed entirely of carbohydrates. The resultant root tissue model focuses on nutrient uptake, transport, and growth, consisting of 149 reactions, 324 genes, and 149 metabolites, while the stem tissue model focuses on transport and growth, consisting of 167 reactions, 291 genes, and 154 metabolites. We have included the root and stem models in the GitHub p-ath773 repository (DOI: 10.5281/zenodo.3735103).

4.6.2.4. Confidence scoring

We have defined reaction confidence scores in a manner consistent with a previously published protocol (Thiele & Palsson, 2010). Confidence scores are integer values between 0 and 4, with higher values corresponding to higher confidence in the inclusion for a given reaction. In the scoring system used, 0 corresponds to an unevaluated reaction; 1 to a reaction included for modeling necessity; 2 to evidence from physiology or a genome annotation; 3 from knock-in knock-out *in vivo* experiments; and 4 for direct biochemical data giving evidence for that metabolic function. The distribution of confidence scores in the component tissue models of p-ath773 can be found in Figures 4.2B through 4.2E. As is shown in these figures, score 3 evidence was not used as score 2 evidence was considered sufficient for model reconstruction and, if greater confidence was required, direct biochemical data could be found since *Arabidopsis* is a model system. Additional information on confidence scoring of the p-ath773 model can be found in Text S1.

4.6.2.5. Curation of these four tissue models

All reactions in all four models have been balanced both in terms of elements and charge. Thermodynamically infeasible cycles have also been resolved by removing reactions, breaking composite reactions, and adding metabolic costs to transport reactions. For all four tissue models, GPR links have been established through a largely automated workflow utilizing the KEGG API (Kanehisa et al., 2017) for the majority of reactions using the code included in the GitHub p-ath773 repository (DOI: 10.5281/zenodo.3735103). This has been followed by having manually curated the GPR links and/or inclusion rational of reactions with non-KEGG identifiers. This information can be found in Data S1. The count of tissue model reactions present in KEGG-defined pathways is shown in Figure 4.2A, giving an overview of each tissue models' metabolic capabilities. The code developed to create these figures is included in the GitHub p-ath773 repository (DOI: 10.5281/zenodo.3735103). The results of this automated workflow can be found in Data S2. Sources for reactions included in leaf, root, seed, and stem models are shown in Figure 4.4B through 4.4E, respectively through confidence scoring (see Text S1).

4.6.3. Linking Tissue Models Utilizing Metabolic Constraints and ORKA

4.6.3.1. Application of ORKA to the p-ath773 model

The application of ORKA to the p-ath773 model is complicated by the fact that there is not a single biomass reaction, but rather four separate reactions, one for each tissue modeled: leaf, root, seed, and stem. Therefore, for the mass of the whole plant, the basic differential equation which defines the change in system mass with time is stated below.

$$\frac{dY_t}{dt} = f(t, Y_t) = \mu_{plant,t} Y_t \qquad (29)$$

Where here Y_t indicates whole-plant mass. This could require some complex hand calculations to determine the value of the RHS of equation (4.29) since $\mu_{plant,t}$ is not calculated in the p-ath773 model, instead only individual tissue biomass growth rates are determined. This leads to a branching point in how to apply the ORKA method to the p-ath773 model: whether to use whole plant mass or individual tissue masses as the basis of biomass calculations. On one hand, as already stated, the biomass of the whole-plant system could be tracked, which would result in a more complex RHS and formulation of $f(t, Y_t)$. On the other hand, the biomass of each plant tissue can be tracked individually as stated in the following equation.

$$\frac{dY_{\theta,t}}{dt} = \mu_{\theta,t} Y_{\theta,t} \qquad \theta \in \Theta \qquad (4.30)$$

It has been decided to use the former method of tracking biomass because solving equation (30) at $Y_{\theta,0} = 0$ yeilds only 0 as a solution. This can be shown in that the generic solution to equation (30) is formulated as follows.

$$Y_{\theta,t+\Delta t} = Y_{\theta,t} e^{\mu_{\theta,t} t} \qquad \theta \in \Theta \qquad (4.31)$$

By the multiplicative identity rule, $Y_{\theta,t+\Delta t} = 0$ if and only if $Y_{\theta,t} = 0$ since the no exponential function can take the value of 0. This presents two issues if this is the method of advancing tissue biomass: i) no tissue can either appear in the system that is not there from the beginning, and ii) no tissue can be removed from the system. This is particularly problematic for a plant system since certain tissues appear and are removed after the plant reaches certain levels of maturity, perhaps most notably flowers and seeds. Therefore, while more complex, determining the value of the right-hand side of equation (4.29) is preferable. Text S1 details the calculation of the RHS of equation (4.20). The end-result of this calculation is as follows.

$$\frac{dY_t}{dt} = \frac{e^{\mu_{leaf,t}} Y_{leaf,t}}{x_{leaf,t}} \left[x_{leaf,t}\mu_{leaf,t} + \frac{d\mu_{leaf,t}}{dt} + \xi_t \right] \tag{4.32}$$

$$\begin{aligned} \xi_t &= x_{root,t}\mu_{root,t} + x_{seed,t}\mu_{seed,t} + x_{stem,t}\mu_{stem,t} \\ &\quad + x_{leaf,t}\left(\mu_{root,t}\zeta_t + \mu_{seed,t}\rho_t + \mu_{stem,t}\delta_t\right)\frac{ds_t}{dt} \\ &\quad + x_{root,t}\frac{d}{dt}(\ln(\omega_t)) + x_{seed,t}\frac{d}{dt}\left(ln(\eta_t)\right) \\ &\quad + x_{stem,t}\frac{d}{dt}(\ln(\lambda_t)) \end{aligned} \tag{4.33}$$

$$\omega_t = \frac{x_{root,t}M_{leaf,0}}{x_{leaf,t}M_{root,0}} \tag{4.34}$$

$$\eta_t = \frac{x_{seed,t}M_{leaf,0}}{x_{leaf,t}M_{seed,0}} \tag{4.35}$$

$$\lambda_t = \frac{x_{stem,t}M_{leaf,0}}{x_{leaf,t}M_{stem,0}} \tag{4.36}$$

$$\zeta_t = \frac{c_{root}\left(c_{leaf}s_t + x_{leaf,0}\right) - c_{leaf}\left(c_{root}s_t + x_{root,0}\right)}{c_{leaf}^2 s_t^2 + 2c_{leaf}s_t x_{leaf,0} + x_{leaf,0}^2} \tag{4.37}$$

$$\rho_t = \frac{c_{seed}\left(c_{leaf}s_t + x_{leaf,0}\right) - c_{leaf}\left(c_{seed}s_t\right)}{c_{leaf}^2 s_t^2 + 2c_{leaf}s_t x_{leaf,0} + x_{leaf,0}^2} \tag{4.38}$$

$$\delta_t = \frac{c_{stem}\left(c_{leaf}s + x_{leaf,0}\right) - c_{leaf}\left(c_{stem}s + x_{stem,0}\right)}{c_{leaf}^2 s^2 + 2c_{leaf}s x_{leaf,0} + x_{leaf,0}^2} \tag{4.39}$$

Note that the above solution makes explicit use of the assumption that the time step used is one hour and that the growth rate unit is inverse hour. Further, an effort has been made in the above formulation to minimize the number of variables used and only the growth rate of the leaf tissue has been included (as the other growth rates of other tissues can be readily calculated from that of the leaf). The number of parameters has not been minimized since readability and compactness have been considered more important than minimizing the number of equations or parameters. This framework requires estimates of time derivatives for some parameters such as seeding level (s_t) and several other quantities in equation (4.33). For all derivative estimates needed, a second-order accurate backwards finite-difference method has been used, as solutions to points previous in time will be known while solutions to points forward in time are unknown. For all parameters for which a derivative estimate needs be made, we have used the following equation where ϕ stands in for any parameter above.

$$\frac{d\phi_t}{dt} \approx \frac{3\phi_t - 4\phi_{t-h} + \phi_{t-2h}}{2h} + O(h^2) \tag{4.40}$$

Note that the trapezoid rule estimates also rely on even step sizes. This allows for smaller error in these estimates. When calculating the derivatives for the first two time points, it is assumed that the parameter values are the same as that for the first time point, e.g. it is assumed that $\phi_{-2h} = \phi_{-h} = \phi_0$. This derivative estimate, particularly the equidistant points requirement, is an important consideration in choosing the particular Runge-Kutta method used in this work. These calculations can be found in the “calculate parameters needed to solve next step equation” lines of the pseudocode in Figure 4.2A.

Given that the steps taken for the Runge-Kutta method have equally spaced values of c_n and that $O(h^2)$ where $h = (c_2 - c_1)\Delta t$ is the order of error for the estimation of the backward derivative, this limits the order of the Runge-Kutta methods which can be chosen for increased accuracy benefits. For instance, a third order Runge-Kutta method has a global error of $O(\mathrm{h}^3)$, less than that of the backward derivative estimate. In this case, choosing any Runge-Kutta method which is higher than third order would merely add complexity with no benefits in terms of the error of the solution. Therefore, a third order Runge-Kutta method has been chosen for implementation with the p-ath773 model. Two commonly-used such methods are Heun's and Kutta's third order rules shown in Figure 4.2B. We have chosen Heun's third order rule for this application as it provides greater accuracy in the integral estimates (e.g. $h = 1/3 \cdot \Delta t$ as opposed to $h = 1/2 \cdot \Delta t$) and there are no negative values in the matrix of parameter a_{ij} of the Butcher tableau which have caused errors in earlier implementations of Kutta's third-order rule to the p-ath773 model (but would not in the current model).

Given that the limiting order of error in this system is the error in parameter derivative estimates, this will be the order of error for ORKA calculations for the p-ath773 model. As the value of h is one third of Δt, we can state that the error would be on the order of $O(1/9 \cdot \Delta t^2)$. This is a significant improvement in the error over previously implementation of dFBA on the *Arabidopsis* models including that of Shaw & Cheung (2018), which by reason of using SOA and a time step of one day (as opposed to one hour) results in much higher error potential. Calculating on the basis of hours the big O error ratio between ORKA and the SOA used by Shaw & Cheung, (2018) the is approximately $1{:}\,5184$. This will provide two distinct advantages to the p-ath773 model. First, higher accuracy for calculations due to smaller step size and lower errors associated with approximations used. Second, increased solution stability, so that more solution steps may be taken without ballooning error.

4.6.4. Other Constraints in the p-ath773's ORKA Framework

The tissue models were linked using techniques similar to a well-known computational framework known for modeling microbial communities (Zomorrodi & Maranas, 2012). This involved specifying how metabolites are allowed to move between tissues in logical ways, which will be described in greater detail later in this section. This framework includes a whole-plant objective which specifies fluxes in each tissue to maximize or minimize. Next, literature information including embryo mass (Hendrik Poorte & Nagel, 2000), initial tissue masses (Baud et al., 2002), growth stages (Boyes et al., 2001), time points at which growth stages occur (Boyes et al., 2001), constraints to link tissue growth rates to appropriate tissue ratios, transpiration (Shipley & Vu, 2017; Sengupta & Majumder, 2014; Schulze, 1986), leaf surface area (Sengupta & Majumder, 2014), usability of provided light (Clauss & Aarssen, 1994; Shipley & Vu, 2017; Juenger et al., 2005), and defining changes in tissue mass ratios (Boyes et al., 2001; Sengupta & Majumder, 2014) has been integrated into these models, which are typically overlooked in most other SMs. In this work, we have decided to simulate *Arabidopsis* biomass across 61 days (1464 hours) of growth, as all plant seeds are dispersed by approximately day 61, and after which *in vivo* data on plant growth and mass is sparse (Boyes et al., 2001). More specific details can be found in the following sub-sections. The full optimization-based framework used in this work has been provided in the GitHub p-ath773 repository (DOI: 10.5281/zenodo.3735103) associated with this work.

4.6.4.1. Enforcing Mass Balance

As concentration is tracked using the ORKA method, the mass balance for the system need not be a strict equality, but rather metabolites should be allowed to be stored and that store should be allowed to be used up. To this end, the mass balance has been defined as follows:

$$\sum_{j \in J} S_{ij} v_j \geq -z_{i,t_n} \qquad \forall i \in I' \qquad (4.41)$$

$$\sum_{j \in J} S_{ij} v_j \geq 0 \qquad \forall i \in I - I' \qquad (4.42)$$

These equations ensure that metabolites might be stored (i.e., more metabolite is produced than consumed) in all cases or available metabolite concentrations might be utilized. When such a concentration is utilized, the LHS of equation (4.41) becomes negative. Hence, the metabolite is consumed at a greater rate than it is produced, yet still obeys the law of concentration of mass by using up the present stores of that metabolite to account for the difference. These equations do not, however, guarantees against the model producing infeasible reaction rates, as large rates of metabolite production are allowable under equation (4.41). To limit the amount of any metabolite stored at a given time point, the following constraint is implemented.

$$\sum_{j \in J} S_{ij} v_j \leq 10 \qquad \forall i \in I' \qquad (4.43)$$

This limits the rate of any metabolite's storage to 10 mmol per gDW tissue per hour and represents that maximum of the allowable violation of the mass balance in the direction of metabolite storage. Such an allowable violation is allowed in all dFBA models where changes in metabolite concentration are allowed. This value is an arbitrary limit on the rate of allowed metabolite storage per hour since it seemed logical to create some limit on metabolite storage rates. While each metabolite likely has its own individual rate, this rate is not reported in literature for many metabolites, therefore a universal, arbitrary number was chosen.

4.6.4.2. Constraints on Metabolite Flow

As has already been suggested, it is not necessarily logical for metabolites to flow between tissues without suitable constraints on that flow. For instance, water is taken up by roots and transported first to the shoot, then to the leaves and seed tissue (if present). It would not, for instance, make sense for water to travel directly from the root tissue to the seed tissue. Therefore, instead of a single metabolite pool connecting tissue, there are metabolite pools connecting each pair of tissues. These pools are shown in Figures 4.3 and 4.5 as arrows and circles between tissues. The following subsections describe how individual metabolites or groups of metabolites are constrained to logical flow through the system.

4.6.4.2.1. Water

Mathematically, for the flow of water these logical metabolite links take the following form.

$$0 \leq -Y_{root,t} v_{root,water_{in}} \leq v_{root,water_{in}}^{bound} \tag{4.44}$$

$$Y_{root,t} v_{root,water_{out}} = -Y_{stem,t} v_{stem,water_{in}} \tag{4.45}$$

$$\begin{aligned} & Y_{stem,t} v_{stem,water_{out}} \\ & = -\big(Y_{leaf,t} v_{leaf,water_{in}} \\ & + Y_{seed,t} v_{,seed,water_{in}}\big) \end{aligned} \tag{4.46}$$

$$v_{leaf,transpiration} = \tau_{leaf}\ (when\ light) \tag{4.47}$$

Equation (4.44) limits the rate of water uptake by the roots to between zero and some pre-defined bound (in the p-ath773 model uptake is defined as a negative flux rate, while output is defined as positive). Equation (4.45) states that all water output by the root goes to the stem and to no other tissue. Equation (4.46) ensures in turn that water output by the stem is taken up by either the leaf or seed tissues. The signs in equation (4.44) through (4.46) ensure consistency with the sign definition of uptake and output in the p-ath773 model. Equation (4.47) enforces transpiration from the plant at a certain level calculated from literature sources (Shipley & Vu, 2002; Sengupta & Majumder, 2014; Schulze, 1986). Transpiration is only allowed during the day because it is assumed that the stomas are open during the day (or when light is available), allowing transpiration, and closed at night (or when light is not available). Transpiration is reported as approximately 2.95 mmol water per m^2 per second (Schulze, 1986). This can be converted to 422.3 mmol water per gDW plant per hour based on the information such as the leaf area ratio (Shipley & Vu, 2002; Sengupta & Majumder, 2014), which is scaled at each time point as appropriate to give the rate in mmol water per gDW leaf per hour.

4.6.4.2.2. Micronutrients

Micronutrients, such as nitrates, sulfates, and phosphates, follow much of the same flow pattern through the plant as does water. This is because water transports dissolved micronutrients to the rest of the plant through the xylem. The major differences are: i) micronutrients will be used up in each tissue so that the amount of each micronutrient leaving each tissue will be less than that entering, which is modeled by equations (4.48) and (4.52) below, and ii) there is no equivalence of transpiration for micronutrients.

$$0 \leq Y_{root} v_{root,\kappa_{in}} \leq v_{root,\kappa_{in}}^{bound} \qquad \forall \kappa \in \mathrm{K} \qquad (4.48)$$

$$Y_{root,t} v_{root,\kappa_{out}} \leq -Y_{root,t} v_{root,\kappa_{in}} \qquad \forall \kappa \in \mathrm{K} \qquad (4.49)$$

$$Y_{root,t} v_{root,\kappa_{out}} = -Y_{stem,t} v_{stem,\kappa_{in}} \qquad \forall \kappa \in \mathrm{K} \qquad (4.50)$$

$$Y_{stem,t} v_{stem,\kappa_{out}} \leq -Y_{stem,t} v_{stem,\kappa_{in}} \qquad \forall \kappa \in \mathrm{K} \qquad (4.51)$$

$$Y_{stem,t} v_{stem,\kappa_{out}} = -\left(Y_{leaf,t} v_{leaf,\kappa_{in}} + Y_{seed,t} v_{seed,\kappa_{in}}\right) \qquad \forall \kappa \in \mathrm{K} \qquad (4.52)$$

Again, signs in the above equations are due to the model convention of denoting uptake of a metabolite as a negative flux, while output of a metabolite is denoted as a positive flux.

4.6.4.2.3. Sucrose

As is well known, sugars in plants are synthesized in photosynthetic tissue, and are transported to the rest of the tissues through the phloem. In the p-ath773 model, it is assumed that the vast majority of photosynthesis occurs in the leaf tissue, and the photosynthetic output of other tissues is negligible. This assumption is based on two factors: i) leaves are tissues specifically designed to carry out photosynthesis, and ii) photosynthesis relies on above-ground surface area to absorb light to drive the process, and leaves have by far the most surface area. Therefore, the flow of sucrose in the modeled plant system is as being exported by the leaf tissue in equation (4.53), transported through the stem tissue in equation (4.54) (allowing for some use of the sucrose by the tissue), and transported to the seed and root via the stem in equation (4.55). These equations are shown below.

$$-Y_{stem,t}v_{stem,sucrose_{in}} = Y_{leaf,t}v_{leaf,sucrose_{out}} \tag{4.53}$$

$$Y_{stem,t}v_{stem,sucrose_{out}} \leq -Y_{stem,t}v_{stem,sucrose_{in}} \tag{4.54}$$

$$\begin{aligned} &Y_{stem,t}v_{stem,sucrose_{out}} \\ &\qquad = -\big(Y_{root,t}v_{root,sucrose_{in}} \\ &\qquad + Y_{seed,t}v_{seed,sucrose_{in}}\big) \end{aligned} \tag{4.55}$$

4.6.4.2.4. Amino Acids

The logical flow of amino acids has been defined explicitly via equations (4.56) through (4.58) stated below, as having been synthesized in the leaf tissue and exported to seed tissue.

$$-Y_{stem,t}v_{stem,x_{in}} = Y_{leaf,t}v_{leaf,x_{out}} \qquad \forall x \in X \qquad (4.56)$$

$$Y_{stem,t}v_{stem,x_{out}} = -Y_{stem,t}v_{stem,x_{in}} \qquad \forall x \in X \qquad (4.57)$$

$$Y_{stem,t}v_{stem,x_{out}} = -Y_{seed,t}v_{seed,x_{in}} \qquad \forall x \in X \qquad (4.58)$$

This is because seed tissue has not been shown to produce all needed amino acids (Lonien & Schwender, 2009), and the root and stem models do not require amino acids for biomass production in the defined biomass composition (Johnson et al., 2007). Essentially, these constraints ensure that all amino acids exported by the leaf are uptaken by the stem, equation (4.56); that these amino acids are not stored in the stem, equation (4.57); and that all amino acids are exported by the stem to the seed tissue, equation (4.58).

4.6.4.2.5. Oxygen and Carbon Dioxide

It is well known that photosynthesis produces molecular oxygen and that respiration produces carbon dioxide. Both processes occur in plants, with photosynthesis necessarily dominating when light is available and respiration dominating when light is not available. As such, in this framework it is specified that the p-ath773 model is a net oxygen producer and net carbon dioxide consumer when light is available whereas the p-ath773 model is a net oxygen consumer and net oxygen when light is not available. These restrictions are formulated in the following equations, where equations (4.59) and (4.60) deal with conditions when light is available for growth while (4.61) and (4.62) apply when no light is available.

When light is available

$$-Y_{leaf,t}v_{leaf,CO_2,in} \geq Y_{root,t}v_{root,CO_2out} + Y_{seed,t}v_{seed,CO_2out} + Y_{stem,t}v_{stem,CO_2out} \tag{4.59}$$

$$-Y_{leaf,t}v_{leaf,O_2out} \geq \left(Y_{root,t}v_{root,O_2in} + Y_{seed,t}v_{seed,O_2in} + Y_{stem,t}v_{stem,O_2in}\right) \tag{4.60}$$

When light is not available

$$0 \leq Y_{leaf,t}v_{leaf,CO_2out} + Y_{root,t}v_{root,CO_2out} + Y_{seed,t}v_{seed,CO_2out} + Y_{stem,t}v_{stem,CO_2out} \tag{4.61}$$

$$0 \leq -\left(Y_{leaf,t}v_{leaf,O_2in} + Y_{root,t}v_{root,O_2in} + Y_{seed,t}v_{seed,O_2in} + Y_{stem,t}v_{stem,O_2in}\right) \tag{4.62}$$

To enforce these constraints, a parameter, called π_t, is defined which takes a value of 1 if light is available for growth and zero otherwise for time t. This is incorporated into the model to simplify the above equations into two equations. Note that the "in" and "out" reactions are combined such that if the model is taking up a given metabolite the reaction rate will be negative, while exporting a given reaction would correspond to a positive reaction rate.

$$\left(Y_{leaf,t}v_{leaf,CO_2} + Y_{root,t}v_{root,CO_2} + Y_{seed,t}v_{seed,CO_2} + Y_{stem,t}v_{stem,CO_2}\right)(1 - 2\pi_t) \geq 0 \tag{4.63}$$

$$\left(Y_{leaf,t}v_{leaf,O_2} + Y_{root,t}v_{root,O_2} + Y_{seed,t}v_{seed,O_2} + Y_{stem,t}v_{stem,O_2}\right)(1 - 2\pi_t) \leq 0 \tag{4.64}$$

The $(1 - 2\pi_t)$ term in the above equations serves as a binary switch alternating between values of -1 and 1 based on the availability of light. In addition to these constraints, some limit must be placed on the uptake of carbon dioxide and oxygen by the leaves of the plant. It has already been noted here the modeled transpiration occurs at 422.3 mmol water per gDW plant per hour (Shipley & Vu, 2002; Sengupta & Majumder, 2014; Schulze, 1986), and this is used as the basis for the exchange of other gasses as well. It is noted that the rate of carbon dioxide uptake is two order of magnitude less than the rate of water loss (Li et al., 1998; Leymarie, Lasceve, & Vavasseur, 1998), and an *in vivo* study identifies the rate of carbon dioxide flow into the leaf as 8 μmol/m^2·s (Regulation & Major, 2007), which is converted using the Leaf Area Ratio (Sengupta & Majumder, 2014) to 1.14 mmol per gDW plant per hour. Assuming standard atmospheric composition (0.04% Carbon Dioxide and 21% Oxygen), then there are approximately 525 oxygen molecules per carbon dioxide molecule at ground level. Here, the limit of oxygen uptake is proportional (in terms of the composition of the atmosphere) to the limit of carbon dioxide uptake, specifically 598.5 mmol per gDW plant per hour is the oxygen uptake limit used. Further, as plants lack a system to transport gasses from one organ or tissue to another (i.e. a circulatory system in the animal sense) it has been assumed that each tissue is responsible for its own gas exchange. As the leaf is a tissue specifically designed for photosynthesis and gas exchange, it will be assumed that the gas exchange occurring in the leaf is at least one order of magnitude larger than that occurring in the rest of the plant. As the other tissues are modeled as heterotrophic (i.e. not significantly photosynthetic), the rate of oxygen uptake must be limited. Therefore, the limit of oxygen uptake for root, seed, and stem tissues is set at 59.85 mmol per gDW plant per hour.

4.6.4.3. Diurnal Carbon Storage Patterns

Plants store carbohydrates in leaf and stem tissues in the form of starch (leaf and stem) and sucrose (stem) in a pattern where the rates of storage may be modeled by a sine wave with a period of 24 hours (Juenger et al., 2005). These equations are defined as follows.

$$v_{leaf,starch_{store}} = A_l \sin(f_l(t + b_l)) \tag{4.65}$$

$$v_{stem,starch_{store}} = A_{st,1} \sin\left(s_{sf1}(t + b_{st,1})\right) \tag{4.66}$$

$$v_{stem,sucrose_{store}} = A_{st,2} \sin\left(s_{sf2}(t + b_{st,2})\right) \tag{4.67}$$

The calculations for defining the necessary parameters namely A_l, s$A_{st,1}$, $A_{st,2}$, f_l, s$f_{st,1}$, $f_{st,2}$, b_l, $b_{st,1}$, and $b_{st,2}$ in equations (4.65) through (4.67) can be found in Data S1. In summary, the necessary parameters listed above have been fit to experimental data by minimizing the sum of squared error between the equations (4.65) through (4.67) using Microsoft Excel's solver tool.

4.6.4.4. Linking Tissue Growth Rates

We have discovered while building this model that tissue growth rates must have enforced links between growth rates of tissues in the system for two reasons: i) linking tissue growth rates allows control of the tissue mass ratios so that they may be modeled as they occur in *Arabidopsis* and ii) this prevents the problem of the model preferentially producing the "cheapest" biomass. The rate of biomass production determined by an SM is the growth rate of the biological system being modeled(Orth et al., 2010); therefore, plant mass can be defined as:

$$M_{\theta,t+\Delta t} = M_{\theta,t}e^{\mu_{\theta,t}t} \qquad \forall\theta \in \Theta \tag{4.68}$$

Further, the ratio of the masses to two tissues can be defined with reference to a single tissue, such as leaf, in the following manner:

$$M_{\theta,t} = \frac{x_{\theta,t}}{x_{leaf,t}}M_{leaf,t} \qquad \forall\theta \in \Theta \tag{4.69}$$

By having substituted equation (4.64) into equation (4.63) and simplifying the result (see Text S1), linear equations have been written to constrain biomass production rates of root, seed, and stem tissues with respect to leaf tissue as follows:

$$\mu_{\theta,t} = \ln\left(\frac{x_{\theta,t+\Delta t}M_{leaf,t}}{x_{leaf,t+\Delta t}M_{\theta,t}}\right) + \mu_{leaf,t} \qquad \forall\theta \in \Theta \qquad (4.70)$$

The quantity inside the natural logarithm is already defined for root, seed, and stem tissues as ω_t, η_t, and λ_t, respectively, in equations (4.34) through (4.36). Therefore, the following constraints are used in the ORKA framework.

$$\mu_{root,t} = \ln(\omega_t) + \mu_{leaf,t} \qquad (4.71)$$

$$v_{seed,biomass} = \begin{cases} 0 & M_{seed,t} = 0, s_{t+\Delta t} = 0 \\ \ln(\eta_t) + \mu_{leaf,t} & M_{seed,t} \neq 0, s_{t+\Delta t} \neq 0 \\ \mu_{leaf,t} & M_{seed,t} \neq 0, s_{t+\Delta t} = 0 \\ -\mu_{leaf,t} & M_{seed,t} = 0, s_{t+\Delta t} \neq 0 \end{cases} \qquad (4.72)$$

$$\mu_{stem,t} = \ln(\lambda_t) + \mu_{leaf,t} \qquad (4.73)$$

Equation (4.72) requires further explanation as to why it is not a single function as equations (4.71) and (4.73). For the first condition, if there is no seed mass at the initial time point and no seeding level at the next time point (meaning the next time point should also have no seed mass) then there should be no growth of the seed tissue. The second condition is when there is both seed tissue at the current point and at the next time point; therefore, this function is analogous to equations (4.71) and (4.73). The final two conditions are artifacts of the exponential nature of the growth rates determined by SMs. The third condition deals with the instance when the seed tissue first appears in the p-ath773 model system. This results in the value of $M_{seed,0}$ being zero, resulting in the limit of $\eta_t \to \infty$ as $M_{seed,0} \to 0$. Similarly, as $\eta_t \to \infty$ then $\ln(\eta_t) \to \infty$. As the model cannot capture infinite growth (and that very high rates of growth would likely result in the auto-cannibalism of existing tissues), we have decided to model the growth in this instance as equal to $\mu_{leaf,t}$. while technically not true, this is because it does set an achievable growth rate for the model. Similarly, at the last time point in which seed tissue is part of the system. This results in $x_{seed} = 0$. As the fraction of seed mass in the system approaches zero ($x_{seed} \to 0$), $\eta_t \to 0$ and as this occurs $\ln(\eta_t) \to -\infty$. Again, this is obviously an issue since infinite negative growth would be both unrealistic and would result in an infinite ray in the p-ath773 model solution, effectively preventing the solution. Instead, similar to the previous case, growth rate is fixed to the negative rate of the growth of the leaf tissue, $-\mu_{leaf,t}$. At this point, it is worthwhile to discuss how seed biomass is lost in a non-productive way (i.e., biomass components are not returned to the metabolic model when seed biomass is lost).

4.6.4.4. Modeling the Loss of Seed Tissue to Seed Dispersal

One of the most metabolically costly activities for many species including for *Arabidopsis* is reproduction. The seed contain a large amount of metabolites which may be metabolized by the embryo to sustain it and allow it to grow before it can photosynthesize. These stored metabolites include fatty acids, proteins, and sugars (Baud et al., 2002). Further, *Arabidopsis* plants produce a very large number of seeds, on the order of approximately 28,000 seeds per gram dry weight of vegetative mass (Clauss & Aarssen, 1994). To properly model this metabolic investment, the model must ensure that these costly metabolites from the seeds are not returned to the plant metabolism when seed biomass is lost. To explain how this could happen, generally the biomass reaction consumes metabolites such as amino acids, fatty acids, sugars, and other necessary compounds in its production and in these cases, $\mu_{seed,t} > 0$. Conversely, when seed mass is being lost from the system, $\mu_{seed,t} < 0$, the biomass precursors are produced from the biomass pseudo-metabolite, and without careful constraints, this loss of seed biomass could cause these precursor metabolites which constitute biomass to remobilize (e.g. used in the metabolism for metabolite production elsewhere) into the metabolic model, resulting in the use of these resources which should be lost to the plant. . Essentially, this would simulate a plant consuming the stores of metabolites in its own seeds, rather than releasing those seeds with its stores intact. Instead, to allow modeling of the complete and non-productive loss of seed biomass, an extra equation called “biomass loss” has been defined to be identical to the biomass equation except it does not produce the biomass pseudo-metabolite. This allows the definition of the following constraint which is in effect during the silique ripening growth stage.

$$\mu_{seed,t} = -v_{biomass\ loss,t} \tag{4.74}$$

This ensures that lost biomass is not re-introduced into the plant metabolism but that it is modeled as lost.

4.6.4.5. Defining the usage of seed stores by the seedling

For the earliest stages of *Arabidopsis* growth, here named as seed germination stage and seed germination to leaf development transition, a seedling’s primary source of carbon is its reserves of stored carbohydrates, proteins, and lipids. It has been shown that seeds have stores of approximately 0.425 μg of sucrose, 6 μg of fatty acids, and 6 μg of proteins (modeled here as component amino acids) available (Baud et al., 2002). As no information concerning the pattern of usage of the seed storage has been found, it has been assumed that the stores are utilized at a constant rate during the duration of the seed germination period and that all the storage is fully consumed by the end of the seed germination to leaf development transition stage, which has been defined the point at which the cotyledons are fully open and leaf development intensifies (Boyes et al., 2001). The rate at which the seedling should uptake the seed storage has been determined by identifying the moles (mmol) of each major component of the seed storage and dividing by the time over which the seedling consumes those. This has resulted in a mmol·h^{-1} quantity. See Data S1 for this calculation. This quantity has then been scaled by plant mass to result in a mmol·gDW^{-1}·h^{-1} quantity, which is used to bound the uptake rates of stored metabolites in the seed. As the leaf has proven to be the most metabolically active tissue, it is assumed that the leaf tissue of an *Arabidopsis* seedling uptakes the stored fatty acids, amino acids, and carbohydrates that are provided for seedling growth during the seed germination stage when the leaves have no access to light (see Figure 4.5, Seed Germination).

4.6.4.6. Defining initial plant and tissue ratios

As the model advances plant and tissue masses with respect to time, the establishment of initial mass for plant and tissues has become important in this framework. Experimental evidence has shown that *Arabidopsis* seeds have a fresh weight (FW) of 25.3 μg and have only about 7% water content (Baud et al., 2002). The embryo itself is assumed equal to the seed mass less the mass of seed stores of sucrose (0.425 μg), Fatty Acids (6 μg), and proteins (6 μg) (Baud et al., 2002). Having assumed that the dry matter content ratio holds for the embryo as well, this has left approximately 11.0 μg dry weight (DW) for the embryo. As information on the ratio of tissue masses in *Arabidopsis* has not been documented in literature, the general ratio for herbaceous plants has been used as a starting point, namely 0.46:0.24:0.3 leaf:root:stem FW (Oakenfull & Davis, 2017). This ratio has been converted to DW ratio for stoichiometric modeling. Experimental data has shown that the dry matter content of leaf tissue is 0.212 DW/FW, of root tissue is 0.170 DW/FW, and of the stem tissue is 0.176 DW/FW (Oakenfull & Davis, 2017). Having converted the FW ratios to DW ratios has given the ratio of 0.511:0.267:0.211 leaf:root:stem DW. While the dry matter content of an embryonic *Arabidopsis* is much higher than that of a mature plant (the source of the utilized dry matter content ratios), this DW tissue ratio has non-the-less been assumed to be accurate for the embryo due to lack of evidence to the contrary.

4.6.4.7. Defining stage times

Time points which define the transition between different stages of growth have been taken from a single source of experimental evidence (Boyes et al., 2001). Stage transitions selected include the transition to stage 0.70 (Seed Germination to Leaf Development transition in Figure 4.5), stage 6.00 (Leaf Development to Flower Production transition in Figure 4.5), and stage 8.00 (Flower Production to Silique Ripening transition in Figure 4.5). Not all lifecycle stage transitions for which there is experimental evidence have been incorporated into this model. In some cases, this has been due to a lack of metabolic relevance, such as the transition from stage 1.04 to stage 1.05 where the plant transitions from 4 rosette leaves to 5 rosette leaves that are greater than 1mm in length. This has not been important to the p-ath773 model as a ratio of plant mass to leaf surface area ratio is used instead[35] (see Data S1). Others that cannot be modeled by the current framework include tissues such as stage 5.10 which is when the first flower bud is visible (Boyes et al., 2001), as the current p-ath773 model has no flower bud tissue. The length of the seed ripening stage has also been determined by experimental evidence (Shipley & Vu, 2002).

4.6.4.8. Defining the change in tissue mass ratios with growth stage

Using available literature evidence, two endpoints for the plant tissue mass ratios have been defined when no seeds are present and all seeds are produced (Boyes et al., 2001). The transition between these states are assumed to be linear with respect to a parameter called seeding, defined above as s. These relationships are then modeled as:

$$x_{leaf,t} = c_{leaf} * s_t + x_{leaf,0} \tag{4.75}$$

$$x_{root,t} = c_{root} * s_t + x_{root,0} \tag{4.76}$$

$$x_{seed,t} = c_{seed} * s_t + x_{seed,0} \tag{4.77}$$

$$x_{stem,t} = c_{stem} * s_t + x_{stem,0} \tag{4.78}$$

$$c_{leaf} = -0.2514; c_{root} = -0.02862; c_{seed} = 0.2030; c_{stem} = 0.07698$$

$$x_{leaf,0} = 0.511;\ x_{root,0} = 0.267;\ x_{seed,0} = 0;\ x_{stem,0} = 0.211$$

Where x_{tissue} has been defined as the tissue mass fraction with respect to the total mass of the plant, c_{tissue} is defined as the change in tissue mass fraction with respect to seeding, and x_{tissue} is defined as the initial mass fraction of each tissue. The gain in the seeding parameter has been assumed to be linear with time and is fit to experimental time point describing the fraction of flowers produced (Boyes et al., 2001) (see Data S1 and Text S1).

4.6.4.9. Defining the availability of light

The amount of light available to the model to use for photosynthesis has been defined initially by literature sources used for other constraints (Oakenfull & Davis, 2017), and scaled by the transmittance of that light source (fluorescent lights) (Baleja et al., 2015) and the absorbance of *Arabidopsis* leaves (Solovchenko & Merzlyak, 2008) and surface area to plant mass of *Arabidopsis leaves* (Li et al., 1998). This has been approximately estimated to be 4.00 mmol·gDW plant^{-1}·h^{-1}. This value has been shown to be 21.50% of the total photons output by the fluorescent light (see Data S1 and Text S1).

4.6.4.10. Defining model maintenance and senescence costs

An important consideration in any SM is the definition of a maintenance cost, which is typically defined as ATP hydrolysis (Thiele & Palsson, 2010). Biomass-based maintenance and senescence costs have been defined as they have been suggested as more accurate or applicable for plant systems (J. H. M. Thornley & Cannell, 1999; Cannell & Thornley, 1999), but have not yet been used in an SM. We have defined maintenance and senescence costs as a biomass drain on each tissue scaled by tissue mass in equation (30). A maintenance cost value of k_m=0.03 day^{-1} has been defined which is in an order of magnitude typical for plant systems (Jeffery S. Amthor, 1984), and the same value has been defined for plant senescence, k_s, as this parameter appears to be generally of the same order of magnitude (J. H. M. Thornley & Cannell, 1999; Jeffery S. Amthor, 1984). These rates are then converted into their per hour equivalent and scaled by tissue mass to enforce these constraints. Only a single constraint has been defined for both phenomena as both are biomass drains whose effect is additive. Literature evidence, including pictorial evidence of plant phenotype at various growth stages, appears to suggest that the rate of plant senescence increases drastically as the flowering production stage finishes and the silique ripening phases begin (in literature, growth stage 0.65 to 9.70) (Boyes et al., 2001). Further, it appears that the plant no longer maintains current mass, but allows tissues to die and desiccate (Boyes et al., 2001). This has been included in the p-ath773 model in that plant senescence is increased by an order of magnitude and plant maintenance is set to zero following the end of the Flower Production stage.

4.6.4.11. Defining model objective functions

For all analyses and results, the objective function of p-ath773 has been to maximize the sum of the biomass production rates for all four tissues according to the following equation (referred to as the default objective).

$$maximize\ z = v_{growth\ leaf} + v_{growth\ root} + v_{growth\ seed} + v_{growth\ stem} \tag{4.79}$$

Where z has been defined as the objective variable with $v_{growth\ tissue}$ being defined as the rate of biomass production, in units of h^{-1}, of the tissue referenced. The maximization of this objective function is approximately equivalent to maximizing the growth rate (change in mass per unit time) of the plant as a whole. This objective function, in early model iterations, has led to one major issue, namely how to avoid the model producing only the metabolically “cheapest” tissue which could result in the maximum objective value but is biologically unrealistic. This is addressed by equations (4.23) through (4.28) and will be further discussed later in the methods section.

It has been noted that the maximization of plant biomass has not been the only feasible objective function for plant SM system; for instance, one alternate objective function is the maximization of plant photonic efficiency (Gomes de Oliveira Dal’Molin et al., 2015; Gomes de Oliveira Dal’Molin, Quek, Palfreyman, Brumbley, & Nielsen, 2010). This objective has generally been framed as minimizing the amount of light used by the plant system, given a required growth rate (Gomes de Oliveira Dal’Molin et al., 2015; Gomes de Oliveira Dal’Molin et al., 2010). As the purpose of this paper is to showcase the ORKA method, rather than the p-ath773 model, alternative objective functions have not been implemented but are possible to implement.

Flux Variability Analysis (FVA) has also been performed on the p-ath773 model which uses all previously defined constraints and the previously defined ORKA method. All flux bounds and constraints are the same and the FVA has an objective function defined as follows:

$$maximize\ or\ minimize\ Z = \sigma_j v_j \tag{4.83}$$

Where the FVA model solution has been iterated for each reaction j, and σ_j has been valued at 1 for the current reaction whose maximum and minimum are to be investigated and 0 for all others and is stepped through first maximizing and then minimizing each reaction. Due to restrictions of the time allowed for model solutions, nine points has been selected at which to perform FVA. These points are 1 hour after germination (HAG, seed germination stage, dark), 70 HAG (seed germination to leaf development transition, light), 90 HAG (seed germination to leaf development transition, dark), 177 HAG (leaf development stage, light), 181 HAG (leaf development stage, light), 770 HAG (flower production stage, light), 810 HAG (flower production stage, dark), 1155 HAG (flower production to silique ripening transition, light), 1170 HAG (flower production to silique ripening transition, dark), 1190 HAG (silique ripening stage, dark), 1199 HAG (silique ripening stage, light). These results generally showed narrow ranges for allowable flux rates.

4.6.5. Symbols Used

4.6.5.1. Sets

I: Set of metabolites in a given model, individual elements are indicated by i.

X: set of amino acids which are synthesized by the leaf and exported to other tissues. $X \subset I$

I': Set of metabolites for which concentration is tracked. $I' \subset I$

U: set of micronutrients which the root uptakes from the soil individual elements are u. U $\subset I$

J: Set of reactions in a given model, individual elements are indicated by j.

Θ: Set of tissues in the model

T: set of time points over which the model is solved, individual elements are indicated by t.

N: Runge-Kutta steps of the chosen or generic Runge-Kutta method, with elements denoted by n.

The final step is denoted n_f, therefore $N = [n_0, n_f]$.

a: Index for a generic set over which a summation is performed.

K: Set of micronutrients uptaken by the roots, with elements denoted by κ. K $\subset I$.

4.6.5.2. Variables

$v_{\theta,biomass,t} \equiv \mu_{\theta,t}$: Rate of biomass production in tissue θ at time t.

$v_{\theta,reaction,t}$: Rate of reaction j in tissue θ at time t.

$z_{i,t}$: Concentration of metabolite i at time t.

Y_t: Biomass concentration at time t. This is used both in the general formulation of SOA and ORKA, as well as to indicate the overall plant biomass.

$Y_{\theta,t}$: Biomass concentration of tissue θ at time t.

4.6.5.3. Parameters

Δt: Size of time step taken by the given DFBA method.

$v_{j,t}^{LB}$: Lower bound of rate of reaction j at timepoint t.

$v_{j,t}^{UB}$: Upper bound of rate of reaction j at timepoint t.

S_{ij}: Stoichiometric coefficient of metabolite i in reaction j.

a: Generic number around which a Taylor series expansion is made.

C_1: Generic parameter of undefined value which appears in intermediate steps for solving ODEs.

$\Gamma_{j,t}$: Multiple application Trapazoid rule-based integral estimate of the integral of $v_{j,t}$ from the first to the second time point.

b_n: Parameter associated with the generic Runge-Kutta of the nth order. These parameters are used to combine Runge-Kutta step size estimates to get the final step size estimate.

k_n: nth step size estimate of the dependent variable made by the Runge-Kutta method.

c_n: Step size of the independent variable in the Runge-Kutta method used. Largest index of c_n is $c_{n_{max}}$.

a_{nm}: Parameters associated with generic Runge-Kutta methods which is used to make sub-steps of the independent variable for estimates of k_n.

$x_{\theta,t}$: Mass fraction of the total plant which is accounted for by tissue θ at time t.

$x_{\theta,0}$: Mass fraction of the total plant which is accounted for by tissue θ at time 0 (initial condition).

c_{θ}: The rate of change in tissue θ mass fraction with respect to seeding. Used to have a linear biomass fraction 'slider' based on the maturity of the plant.

s_t: Level of seeding at time t. This parameter is used to indicate plant maturity and to simulate the increase in seed tissue mass fraction (and corresponding decrease of other tissues) as time passes.

ξ_t: Parameter used to split the calculation of $\frac{dY_t}{dt}$ into multiple equations to make the formulation more readable.

ω_t: Parameter used to split the calculation of $\frac{dY_t}{dt}$ into multiple equations to make the formulation more readable. Deals with the change in plant mass fraction that is root tissue.

η_t: Parameter used to split the calculation of $\frac{dY_t}{dt}$ into multiple equations to make the formulation more readable. Deals with the change in plant mass fraction that is seed tissue.

λ_t: Parameter used to split the calculation of $\frac{dY_t}{dt}$ into multiple equations to make the formulation more readable. Deals with the change in plant mass fraction that is stem tissue.

ζ_t: Parameter used to split the calculation of $\frac{dY_t}{dt}$ into multiple equations to make the formulation more readable. Deals with the change in plant mass fraction that is stem tissue.

ρ_t: Parameter used to split the calculation of $\frac{dY_t}{dt}$ into multiple equations to make the formulation more readable. Deals with the change in plant mass fraction that is stem tissue.

δ_t: Parameter used to split the calculation of $\frac{dY_t}{dt}$ into multiple equations to make the formulation more readable. Deals with the change in plant mass fraction that is stem tissue.

ϕ: Generic time-dependent parameter, used to show an equation that applies to a number of parameters.

τ_{leaf}: Transpiration rate of water from the leaf when the stomata are open during the day. Calculated from literature.

$v_{root,water_{in}}^{bound}$: Limit on rate at which the root tissue can take up water.

$v_{root,\kappa_{in}}^{bound}$: Limit on the rate at which the root tissue can take up micronutrients

A_l: Amplitude of diurnal starch storage pattern in leaf.

f_l: Frequency of diurnal starch storage pattern in leaf.

b_l: X-intercept of diurnal starch storage pattern in leaf.

$A_{st,1}$: Amplitude of diurnal starch storage pattern in stem.

$f_{st,1}$: Frequency of diurnal starch storage pattern in stem.

$b_{st,1}$: X-intercept of diurnal starch storage pattern in stem.

$A_{st,2}$: Amplitude of diurnal sucrose storage pattern in stem.

$f_{st,2}$: Frequency of diurnal sucrose storage pattern in stem.

$b_{st,2}$: X-intercept of diurnal sucrose storage pattern in stem.

π_t: Binary parameter whose value states whether or not a

4.6.5.4. Functions

$f(t)$: Generic function dependent on t.

$f(t, Y_t)$: Generic function dependent on t and Y_t.

$f^{(n)}(t)$: n^th derivative of generic function $f(t)$.

$g(t)$: Function estimated from $f(t)$ using a Taylor series expansion.

$O(h^n)$: Big O notation used to indicate the order of error for an estimated function, where h is the variable by which the error is defined an n is the order of that error.

Chapter 5

5. OPTIMIZATION-BASED EUKARYOTIC GENETIC CIRCUIT DESIGN (EUGENECID) AND MODELING (EUGENECIM) TOOLS: COMPUTATIONAL APPROACH TO SYNTHETIC BIOLOGY

Portions of this material have been submitted for publication and are currently under review.

5.1 PREFACE

Synthetic biology has the potential to revolutionize the biotech industry and our everyday lives and is already making an impact. Developing synthetic biology applications requires several steps including design and modeling efforts which may be performed by *in silico* tools. In this work, we have developed two such tools, Eukaryotic Genetic Circuit Design (EuGeneCiD) and Modeling (EuGeneCiM), which use optimization concepts and bioparts including promotors, transcripts, and terminators in designing and modeling genetic circuits. EuGeneCiD and EuGeneCiM preclude problematic designs and could lead to synthetic biology application development pipelines. EuGeneCiD and EuGeneCiM are applied to developing 27 basic logic gates as genetic circuit conceptualizations which respond to heavy metal ions pairs as input signals for *Arabidopsis thaliana*. For each conceptualization, hundreds of potential solutions were designed and modeled. Demonstrating its time-dependence and the importance of including enzyme and transcript degradation in modeling, EuGeneCiM is used to model a repressilator circuit.

5.2 INTRODUCTION

Synthetic biology is the design of living systems, utilizing engineering principles, to accomplish a desired task or purpose (Khalil and Collins, 2010). To date, applications include novel biochemical synthesis pathways and many biological analogs of electronic circuits such as logic gates, sensors, toggles, oscillators, and switches (Khalil and Collins, 2010; Kim and Winfree, 2011; Liu and Stewart, 2015; Scheller *et al.*, 2020) with a long term goal of programmable biology (Xia *et al.*, 2019). Commercial products which are the result of applications of synthetic biology are emerging in restaurants (the Impossible Burger), pharmacies (Januvia indicated for diabetes), electronics (Hyaline used in foldable smartphones), and hospitals (Kymriah, a cell-based therapy indicated for B-cell acute lymphoblastic leukemia) highlighting the emerging roles of synthetic biology throughout society (Voigt, 2020). Therefore, the tools which aid in the development of novel synthetic biology applications will be of both scientific and commercial value to accelerate the development of new applications. There are five major stages in the development of a new synthetic biology application: conceptualization, design, modeling, construction, probing, testing, and validation (Liu and Stewart, 2015). Of interest are the design and modeling stages, which with the proper tools, could be largely automated to create a synthetic biology application development pipeline from conceptualization to construction.

In the design of new applications, synthetic biology often relies on the intuition of biologists and engineers; their knowledge of available promotors, genes, terminators, transcripts, enzymes, and proteins (collectively, bioparts) and the associated systems; and their design ability to create new applications. This approach is generally limited to system experts and to designs which are intuitive. Alternatively, a computational model-driven approach is advantageous in that it allows for non-intuitive designs and the quick *in silico* screening thereof, so that only designs with the greatest chance of success are constructed. Several design and modeling tools exist such as Cello 2.0 (Chen *et al.*, 2020), OptCircuit (Dasika and Maranas, 2008), the work of Zomorrodi and Maranas (2014) (the tool was unnamed), EQuIP (Davidsohn *et al.*, 2015), SynBioSS (Hill *et al.*, 2008), and several others which may be adapted to various systems and to screening of genetic circuits (Liu and Stewart, 2015). Figure 1 summarizes the unique approach to the problem of design along with advantages and disadvantages of each of these tools within the context of developing synthetic biology applications. Although these tools have successfully designed or simulated behaviors replicated *in vivo*, the most overarching challenge associated with these tools is their specialization for design or modeling tasks which has no clear workflow or method by which to link the two activities. This is highlighted in that some design tools, such as Cello 2.0, published synthetic biology workflows which skip the modeling step altogether and used more expensive and time-consuming *in vivo* screening processes (Borujeni *et al.*, 2020). A particularly difficult problem in current optimization-based design tools such as Zomorrodi and Maranas (2014), and OptCircuit (Dasika and Maranas, 2008) are Bistable Orthogonal Designs (BODs). These produced design solutions that would not function as desired. For instance, consider the example shown in Figure 2, where it is desired to produce a circuit with an AND response to copper and zinc ions using a GFP reporter. Using only a handful of parts, it is possible to produce a circuit with two stable states (where both tetR and GFP are produced or only cI is produced). Further, these two stable states are independent of (or orthogonal to) the signals which the circuit should respond (e.g. the copper and zinc ions). For a BOD, a solver might then pick whichever state is necessary to match the desired conceptualized circuit behavior irrespective of the conditions, rendering the circuit effectively useless for the proposed application. These BODs are technically correct solutions to the conventional optimization-based tools but require further manual scrutiny to identify and remove these problematic solutions. When producing large numbers of solutions, BODs generally outnumber true designs and can overwhelm a researcher's ability to screen.

One promising area for synthetic biology applications is in plants, particularly commercially important crops such as maize (*Zea mays*), rice (*Oryza sativa*), and barley (*Hordeum vulgare*). Applications in plants include increasing nutrient content (Beyer *et al.*, 2002; Gonzali, Mazzucato and Perata, 2009), synthesizing novel chemicals (Liu and Stewart, 2015; Mortimer, 2019), improved crop resilience (Pixley *et al.*, 2019), and synthetic sensors (Liu and Stewart, 2015). Here, we have chosen to demonstrate the EuGeneCiD and EuGeneCiM tools using the model plant species *Arabidopsis thaliana* (hereafter, Arabidopsis) because it is well studied and has been used for many synthetic biology applications (Holland and Jez, 2018a). We have further chosen to design and model plant-based synthetic sensors of heavy metal in the root of Arabidopsis. Heavy metal pollution occurs as a result of human activities (such as mining or manufacturing), and is toxic to living organisms at sufficient concentrations, even essential elements such as Zinc. These metal ions can enter the soil via several possible routes including from water and the air (Vardhan, Kumar and Panda, 2019; Vareda, Valente and Durães, 2019). Three of the most common heavy metal pollutants are Copper, Cadmium, and Zinc, (Vardhan, Kumar and Panda, 2019b) to which Arabidopsis has some natural response mechanisms. By creating reporter systems which respond to these heavy metal ions, it may be possible in the future to develop synthetic biology applications in crop species for metal ion removal or mitigation from contaminated soils through phytoremediation (Jacob *et al.*, 2018). Different logical combinations of present ions might require different phytoremediation strategies; therefore, the construction of logic gates responding metal ion signals would be a logical first step in the long-term development of these strategies and applications.

For developing a combined design and modeling workflow, in this work, we developed two optimization-based tools, namely Eukaryotic Genetic Circuit Design (EuGeneCiD) and Modeling (EuGeneCiM), which utilize an input of the conceptualized circuit behavior and perform an automated simulation of the optimal and suboptimal circuit designs for manual screening. EuGeneCiD provides one key improvement upon previous optimization-based tools (Ali R. Zomorrodi and Maranas, 2014; Dasika and Maranas, 2008) by developing constraints (called the attribution constraints) which precludes BODs. In addition, several other distinct differences and improvements distinguish the EuGeneCiD tool from either of these previous works. Firstly, EuGeneCiD is designed for eukaryotic systems where Ribosome Binding Sites (RBSs) are not a critical design element, but replaces such elements with terminators which are important in eukaryotic gene expression, particularly for plants (F. de Felippes *et al.*, 2020; Nagaya *et al.*, 2010). Secondly, the rate of mRNA and protein degradation on circuit behavior are incorporated, which leads to new design possibilities. Third, the tool was made more granular so that concentration values are not always integer values. Fourth, the layers of the central dogma (transcription and translation) are mathematically separated so that, aside from relative concentration levels, relative levels of mRNA for genes might also be designed and simulated. EuGeneCiM takes these unique elements and, utilizing a design passed from EuGeneCiD, simulates circuit behavior over a given number of hypothetical time points, which will allow for screening of circuit behavior before constructing these proposed synthetic biology applications.

Using bioparts, which are either a part of natural Arabidopsis heavy-metal response mechanisms, or shown to function in Arabidopsis from other species, and fluorescent proteins as state reporters, EuGeneCiD is applied to developing these synthetic heavy metal sensors in Arabidopsis. EugeneCiD was used to create design solutions for 27 different genetic circuits formed from combining nine unique two-input logic gates with three different input signal pairs. These input signals are the presence of Cadmium, Copper, and Zinc ions at high or toxic concentrations. For each genetic circuit conceptualization which was able to be designed from the given biopart library, EuGeneCiD generated hundreds of feasible solution designs, each with a corresponding dynamic simulation from EuGeneCiM. Aside from basic logic circuits, repressilators have also proven to be a useful control schema in synthetic biology, allowing for oscillating gene expression (English, Gayet and Collins, 2021). Therefore, EuGeneCiM is used to model the dynamic behavior of a repressilator circuit to demonstrate its utility as a stand-alone dynamic modeling tool and the value of incorporating mRNA and protein degradation in modeling efforts. Together, the EuGeneCiD and EuGeneCiM tools can hypothesize genetic circuit designs and simulate their behavior to increase the chances that a plant might have the desired behavior when transformed, potentially saving time and resources. This work could be the basis for the development of a synthetic biology application pipeline. Therefore, for the ease of use and the facilitation of this pipeline, various programs have been developed to make EuGeneCiD and EuGeneCiM user-friendly. Further, the design solutions produced here could form the basis of future heavy metal phytoremediation applications of synthetic biology particularly in important crops like *Zea mays* (maize). Maize has been identified as both Cadmium tolerant (Rizwan *et al.*, 2017) and as a Cadmium hyperaccumulator (Wuana and Okieimen, 2010), and is already used for heavy metal phytoremediation (Rizwan *et al.*, 2017). Additionally, maize has been identified as a bioaccumulator of both Zinc and Copper (Sekara *et al.*, 2005)(Wuana and Okieimen, 2010). From this, maize is already particularly well suited for phytoremediation applications, and could be engineered through synthetic biology to be superb, solving multiple problems at a stroke by providing food from otherwise toxic farmland while cleansing it of heavy metal ions toxic to both humans and other plants.

5.3. RESULTS

5.3.1. Selection of Test System and Synthetic Biology Conceptualizations

Arabidopsis was chosen as the test system for the development and subsequent application of the EuGeneCiD tool since it is a model plant system to which systems biology has often been applied (Holland & Jez, 2018b). It was decided to develop heavy metal ion biosensors in the Arabidopsis root, which would report sensor state using fluorescent proteins. A plant system, in particular, was chosen for this work because in the future EuGeneCiD and EuGeneCiM will be applied to plants of biotechnological and agronomic importance (e.g., *Zea mays*) for various applications related to plant health and fitness, potentially including phytoremediation of heavy metal pollution. Since phytoremediation strategies may change depending on the metal ion(s) present, basic logic gates are conceptualized here which report on the presence or absence of the metal ions.

5.3.2. Development of the Eukaryotic Genetic Circuit Design (EuGeneCiD) Tool

EuGeneCiD was conceived and developed to address the limitation of the current state-of-the-art optimization-based design tools for synthetic biology applications (Ali R Zomorrodi and Maranas, 2014; Dasika and Maranas, 2008). Particularly, by changing the focus to eukaryotic systems, allowing granularity, modeling transcript abundance, adding terminators as a design element (which are particularly important in plant synthetic biology), and creating the attribution constraints. The initial EuGeneCiD formulation was inspired by other optimization-based circuit design works (Ali R Zomorrodi and Maranas, 2014; Dasika and Maranas, 2008) and was formulated specifically to apply to eukaryotic systems and incorporate biopart degradation. This involved using terminators as opposed to RBSs as part of the design; incorporating mRNA and protein degradation; having a more granular values of concentration; and reporting relative mRNA abundance for particular genes. Attempts were made to incorporate time to make EuGeneCiD a dynamic design tool. This would influence various design variables, such as concentration, yet this proved computationally intractable and was abandoned. At this stage in development, it was decided to separate the formulations of design and modeling tools. When applying this first version of the EuGeneCiD tool to a modest sized biopart database, the issue of BODs became apparent and pressing. The final stages of the development of EuGeneCiD involved the creation of the attribution constraints to prevent BODs. These attribution constraints account for a high fraction of all constraints (about 42%) and variables (about 42% of total) in the formulation of EuGeneCiD and thus account for a fair amount of the tools' computational expense. This tradeoff is considered worthwhile in that it allows for the preclusion of BOD solutions which can account for greater than 90% of solutions in some instances when the attribution equations are not included. The final formulation of EuGeneCiD is a Mixed Integer Linear Programming Problem, with a single-level objective function maximizing the concentration of desired enzymes and minimizing that of undesired enzymes. Initial testing of EuGeneCiD was conducted using hypothetical bioparts, details of which are provided in the GitHub repository associated with this work (github.com/ssbio/EuGeneCiDM or DOI: 10.5281/zenodo.4762590). The final formulation has over three dozen constraints and variables which are detailed in the methods section.

5.3.3. Development of the Eukaryotic Genetic Circuit Modeling (EuGeneCiM) Tool

EuGeneCiM was conceived and developed to address the lack to optimization-based tools for the modeling of proposed synthetic biology application designs, particularly one which might readily be passed designs for screening. As previously stated, EuGeneCiM initial development began when it was noticed that including time-based simulations inside the EuGeneCiD tool was computationally intractable. EuGeneCiM is similar to EuGeneCiD in formulation with three major exceptions. First, the design variable is made a parameter in EuGeneCiM as these values are passed from an optimal or suboptimal solution of EuGeneCiM. Second, as EuGeneCiM does not design, the attribution constraints are unnecessary and therefore unused, thus considerably boosting solution speed. Third, as the design is not variable, this allows certain simplifications in the formulation. Initial testing of EuGeneCiD was conducted using hypothetical bioparts, provided in the GitHub repository associated with this work (github.com/ssbio/EuGeneCiDM). The final formulation has approximately two dozen constraints and variables which are detailed in the methods section.

5.3.4. Definition of the Bioparts Database

Following the creation and initial testing of the EuGeneCiD and EuGeneCiM tools, a database of bioparts was created for the design of genetic circuits which respond to Cadmium, Copper, or Zinc ions, or combinations thereof to design and simulate various logic gates. Note that bioparts which are responsive to the metal ions do not directly respond to those ions, but rather make use of the native metal sensing or signaling pathways of Arabidopsis and are bioparts whose activity is affected by these signaling pathways. This approach is used because it was decided that it would be too complex to introduce the various signal pathways in a target organism with each design. Promotors included in the biopart database are shown in Figure 5.3. Details on the sources for these bioparts, their parameterization, and their reason for inclusion in the database can be found in Supplemental Table S1.

5.3.5. Application of EuGeneCiD and EuGeneCiM

The EuGeneCiD and EuGeneCiM tools are embedded in the workflow shown in Figure 5.4. In summary, this workflow uses the bioparts library and the synthetic biology application conceptualization as inputs from which the EuGeneCiD problem is attempted. Should a solution be found, EuGeneCiM is solved across several time points to model the designed circuit. If a solution is not found, there are two possibilities: all possible designs with the specified parameters (primarily circuit size) have been identified, or that all possible designs have been identified which are smaller than some maximum allowed circuit size. In the former case, the size of the sought design is incremented, and EuGeneCiD is attempted again. Otherwise, the selection of designs is returned, and the user may select a design from the design and modeling information. For greater details, see Methods.

To demonstrate the utility of EuGeneCiD and EuGeneCiM tools, it was decided to use these tools to design and model 27 unique genetic circuit conceptualizations using the defined bioparts database. Each conceptualization will have its own input file, an example is provided in Supplemental Table S2, containing all information from Supplemental Table S1 in addition to a logic table, and a parameter specifying the number of time points to model. These unique conceptualizations were defined both by the logic circuit and the ligand pair to which that circuit is to respond. The logic circuits to which EuGeneCiD and EuGeneCiM were used to design and model include AND, NIMPLY, converse non-implication (abbreviated CNI), HALF ADDER, NAND, NOR, OR, XNOR, and XOR. Note that CNI is included because it is logically equivalent to NIMPLY with a reversed ligand order. Divalent heavy metal ion pairs, representing common heavy metal pollutants (Vardhan et al., 2019b), were selected to serve as the signals for the logic gates by their presence or absence. The metal ion signal pairs used are Cadmium and Copper; Cadmium and Zinc; and Copper and Zinc. The table in Figure 5.5 shows each combination of metal ion pair and logic gate.

It should be noted that the applications of EuGeneCiD and EuGeneCiM presented here do not make full use of the in-built capabilities of these algorithms. First, these algorithms have the potential to consider alternative splicing, through definitions of the variable which maps transcripts to its encoded enzyme (ρ_{je}) and transcriptional efficiency (η_j). The former can be used to define more than one transcript-enzyme encoding relationships and the latter can be lowered to reflect fractions of transcript being used to encode each alternative splice. In addition, the capability exists for enzymes to be regulated by environmental cues and other enzymes. These capabilities are not exploited in this application because it was desired to apply these tools to a plant system, and Arabidopsis appears to not have such sophisticated bioparts natively, nor have such parts been engineered for Arabidopsis. However, these capabilities will function in the event that they are needed and defined in the input bioparts library, as these functions have been tested using test databases.

5.3.6. General EuGeneCiD Solution Trends

Several general trends emerge from the sets of solutions produced by EuGeneCiD and can be identified in Figure 5.5. First, as highlighted in Figure 5.5, using the given database, it appears that certain simpler logic gate such as AND, NIMPLY, NOR, and OR are easier to find design solutions for. This is indicated by high numbers of solutions after the seven day run time, short solution times (minimum, average, and maximum), and a large percentage of reported solutions being proven optimal solutions (as opposed to integer solutions which do not guarantee optimality). On the other hand, circuits such as XNOR, XOR, and HALF ADDER are generally more difficult to find design solutions as indicated by fewer solutions, longer solve time, and low percentage of reported solutions being proven optimal. For these circuits, the majority of solutions are integer solutions without proven local or global optimality. In addition, these more difficult circuits generally also have higher minimum and mode circuit sizes, as well as longer solution times. These circuits are also are more likely to have been terminated by reaching the seven day time limit, as opposed to the easier circuits which were more likely to be terminated by reaching the maximum number of allowed solutions. As shown in Figure 5.5, more complex solutions generally require more triads (size is the number of triads in the design) to achieve the desired logic.

A particularly interesting trend in EuGeneCiD solutions, shown in Supplemental Figure 5.S1, is that the maximum objective function value never occurs in the first solution, with the exception of the Cu^{2+}/Zn^{2+} XOR responsive circuit, though the minimum objective value sometimes occurs at this point. This can be for multiple reasons. The objective function is defined as the difference of response strength under desired response conditions and response strength under undesired condition. This formulation ideally will favor solutions with strong responses and low expression leakiness. See Methods for the mathematical formulation. The first possibility is that a biopart with this inherent function might be leaky or not particularly strong, yet would be the simplest possible solution. A second possibility is, due to the nature of the EuGeneCiD objective function, different circuit conceptualizations will have slightly different priorities in their optimal designs. In summary, depending on the sparsity of the response vector(s) in the input logic table, a slight favoritism for low leakiness of the response protein(s) or for a strong response pulse may be favored. A full discussion of this can be found in the Methods.

5.3.7. Dissecting Selected Circuit Designs

This study produced a very large number of design and modeling results, more than 23,000 to be precise. This volume allows for analysis of the broad solution trends discussed while precluding the analysis of each individual solution. All solutions may be found in the associated GitHub repository (github.com/ssbio/EuGeneCiDM). Additionally, code is provided in the repository which will plot any given solution (see the provided documentation in GitHub). This code was used in part to generate the graphs in Figure 5.6. By investigating several solutions using this code, we have selected three representative circuit results (two of which could be defined as successful and one is unlikely to be) as example results, shown in Figure 5.6. One general feature of interest in the EuGeneCiM tool can be seen in each of the modeling results graphs: the start-up time. EuGeneCiM essentially assumes that the genetic circuit is newly introduced into the target organism at time point 0 therefore, there is some delay (2 time points) between introduction of the circuit and the response of the circuit to environmental conditions. A second point of interest is that, while both tools use Mixed Integer Linear Programming, the curves produced are non-linear. This is because, in EuGeneCiM the half-life based degradation of transcripts and proteins is calculated between time steps as a “carry over” value from one time point to the next (as shown in the workflow image Figure 5.4 and described in the Methods).

The first successful example, solution #41 for a Cd^{2+}/Cu^{2+} responsive AND circuit, is shown in the top third of Figure 5.6. Solution #41 was chosen as it is the Cd^{2+}/Cu^{2+} responsive AND circuit with the maximum objective function value, likely due to the multiple gated encoding of GFP. This solution contains four triads (promotor/gene/terminator groupings which specify the circuit design): P_{FRO2}/gene_cI/HSPt, P_{ara}/gene_cI/CaMV25St, P_{RM}/gene_GFP/HSPt, and $P_{EXO70B1\_11}$/gene_GFP/HSPt. There are two responsive elements to the signal ions, promotors P_{FRO2} (responding to Cd^{2+}) and $P_{EXO70B1\_11}$ (responding to Cu^{2+}). These then regulate the expression of GFP indirectly and directly, respectively. Note that while P_{ara} is regulated by araC, because araC is not encoded, it will act like a constitutive promotor. Due to the short half-life of cI, this circuit maintains a constitutive pool of cI which is below the concentration threshold necessary for a cI-expressing phenotype unless Cd^{2+} is present. This gates the expression of GFP from the P_{RM}/gene_GFP/HSPt triad, preventing GFP expression from this triad unless Cd^{2+} is present. GFP expression induced by Cu^{2+} is regulated directly. This causes the circuit to be quicker to respond to the presence of Cu^{2+} than to Cd^{2+} in the modeling results. The double-encoding of the GFP results in the significantly stronger response of the circuit to both conditions, than to a single condition. This is one potential drawback of the binary encoding of the conceptualization in that there is no mechanism to ensure equal expression in all cases where expression is desired, since phenotype is what is desired, rather than strength of that phenotype.

The second successful example, solution #11 of a Cu^{2+} NIMPLY Zn^{2+} circuit, is shown in the middle third of Figure 5.6 also uses cI as the desired control enzyme which gates expression of GFP. This circuit uses three triads in the design: P_{GSTF1}/gene_cI/HSPt, P_{FDR3}/gene_cI/NOSt, and P_{RM}/gene_GFP/HSPt. For controlling the expression of cI, a moderately strong promotor, P_{FDR3} (which is repressed by Zn^{2+}), is paired with a relatively inefficient terminator NOSt, which results in a pool of cI transcripts which can quickly build or degrade in the absence or presence of Zn^{2+} but which is not sufficient for cI-expression phenotype. The P_{GSTF1}/gene_cI/HSPt triad then is also a deciding factor in cI phenotype, encoding stable RNA (from an efficient terminator, HSPt) from a moderate promotor (P_{GSTF1}). This second promotor results in a slowly building yet stable pool of cI transcripts. When both triads produce cI, the concentration is high enough for cI expression. When cI is expressed, the very strong promotor P_{RM} is activated, resulting in strong GFP expression. When modeled, this mixed approach to cI production (using from quick- and slow- accumulating pools of cI transcript) in combination with the sort half-life of cI results in a slow-responding circuit (only beginning to diverge from other conditions at time point 7), as expression from both triads is required. Yet, when cI is at sufficient concentration, the circuit responds very strongly. It is highly possible that the response strength would be greater than what is shown if the circuit were modeled for more time points. Theoretically, this circuit could be quickly "shut off" by lack of a Cu^{2+} signal or especially the presence of a Zn^{2+} signal. Due to the single-encoded gene_GFP, GFP expression is uniform and low in non-expressive conditions.

The provided unsuccessful solution is solution #26 for a Cd^{2+}/Zn^{2+} responsive NAND circuit, shown in the bottom third of Figure 5.6. As with the previous example, three triads are used, two of which gate the expression of GFP through a control enzyme, in this case, araC. The triads of this design are P_{CdI3}/gene_araC/CaMV25St, P_{HYP1}/gene_araC/NOSt, and P_{ara}/gene_GFP/HSPt. One interesting point to note is that the used promotors are weaker and terminators are less efficient than those generally used with cI because the control enzyme, araC, has a longer half-life. In this unsuccessful example, the circuit responds correctly to the presence of both Cd^{2+} and Zn^{2+}; of Zn^{2+}; and to no signal. This circuit fails in the condition at which only Cd^{2+} is present. This is because, while EuGeneCiD partially accounts for enzyme degradation, it does not account for accumulation as it predicts that under this condition araC will not accumulate sufficiently to be active. However, when accounting for accumulation, EuGeneCiM predicts that araC will accumulate enough for an araC-expressed phenotype around time point 5, resulting in a sharp decline in GFP response from this point. This circuit could be potentially corrected by replacing the terminator in the P_{CdI3}/gene_araC/CaMV25St triad with a less efficient terminator. Unlike the other examples, this also illustrates that the trend of EuGeneCiM models might change direction and even sign during the simulation. This change during the simulation may result in a correct circuit response, whereas previous time points might suggest an incorrect response (consider the condition where both Cd^{2+} and Zn^{2+} are present). This suggests that, for some circuits, it may be useful to look at longer-term behavior in some cases where a designed circuit may be modeled to show an incorrect response.

5.3.8. EuGeneCiM-modeled Repressilator

To demonstrate the utility of EuGeneCiM as an independent modeling tool, it was decided to model a repressilator circuit. Repressilator circuits rely on the degradation of proteins whose expression is repressed to allow a downstream protein to be expressed, and therefore could not be modeled by non-dynamic genetic circuit modeling tools, or tools which do not consider transcript or protein degradation. A five triad repressilator circuit was manually designed (because a repressilator cannot be designed by the non-dynamic EuGeneCiD) and is shown in Figure 5.7. This circuit utilizes araC, cI, and tetR control enzymes from *E. coli*, which have been reported to be used in synthetic biology applications in Arabidopsis (Messing, 1998), are well characterized, and which control promotor expression. All these enzyme inhibit one promotor in the biopart library, and importantly two of these enzymes have corresponding promotors which they activate, araC and cI. No promotor could be found which was activated by tetR. These activated promotors encode reporting fluorescent enzymes mKO (activated by araC) and GFP (activated by cI) identified through the fluorescent protein database (fpbase.org). Using EuGeneCiM, it was decided to model the first 100 relative time points of the simulation of the repressilator.

This simulation highlights several important features of the EuGeneCiM for which there was no opportunity for discussion when modeling EuGeneCiD-created designs. First, transcript production, transcript level (shown in Figure 5.7C), enzyme production, and enzyme level (shown in Figure 5.7B) are all modeled and tracked by EuGeneCiM (complete results can be found in the GitHub associated with this work at github.com/ssbio/EuGeneCiDM). Second, the shape of the response curves is of interest. As shown best by the tetR response curve (purple), EuGeneCiM models can achieve steady state (or near steady state) and be perturbed from that state. This curve also shows that EuGeneCiM is capable of modeling oscillatory circuit designs. This indicates that EuGeneCiM is not wholly dependent on EuGeneCiD and can be used as an independent modeling tool. Further, upon introducing three enzymes, there is some unsteady-state start-up period where the enzymes in question are all produced prior to some control enzyme taking dominance. Using GFP as an example, this period is approximately the times from time points 0 to 12. This is the start-up period, and varies to some extent between enzymes, though it appears that GFP has the longest such period. It can also be seen in these graphs that the amplitude of enzyme responses are uneven between enzymes. This is due to differences in promotor strength, (stronger promotor, higher peak), terminator efficiency (more efficient terminator, higher peak), and enzyme half-life (longer half-life, higher peak). These factors also influence the breadth of the peaks, with shallower peaks also being broader, and taller peaks being narrower, with cI and GFP as the two more extreme cases in each direction, respectively. Though it should be noted that regardless of the breadth or height of the peaks, all enzyme expressions have a period of 22 time points, a period that is indefinitely stable (this repressilator has been modeled out to 500 time points).

One potential discrepancy with *in vivo* behavior is that repressilator responses *in vivo* are generally sinusoidal in behavior, in EGeneCiM models, the behavior is not perfectly sinusoidal in shape with sharp discontinuities at peak and trough. This is because transcription of a triad is modeled as a binary (either transcribed or not), rather than as a more continuous process as might occur *in vivo*. However, this wave has several similarities to a sine wave including a well-defined period (22 time points), amplitude (approximately 8 units), y-intercept (varies depending on the enzyme of interest, for GFP it is 10.37 units, defined from the average post-start-up), and x-intercept (varies depending on the enzyme of interest, for GFP this is 2 units). Despite their slightly different shape, they still are quite similar to sine waves nonetheless. As a demonstration of the modeled GFP enzyme level's similarity to a sine wave, a sine wave with the aforementioned characteristics of the GFP expression curve, graphs are provided in Supplemental Table S3 which highlight the similarity of the GFP enzyme level curve shape and that of a sine wave. This has also been done for cI. The Pearson correlations between these curves are r=0.91 and r=0.97, respectively, showing a strong linear relationship between the curves produced by EuGeneCiM and the sine waves produced by using the characteristics of those curves, suggesting that the shape of the curves are very similar. Further, these curves have the same mean value (about 10.4 units), and similar standard deviations (5.7 units for the sine wave and 6.0 units for the GFP curve) suggesting very similar magnitude, in addition to similar shape.

5.4. DISCUSSION

Synthetic biology holds great potential for technological advancements and applications in a wide variety of fields. The designing of a new application involves five distinct steps, of which the first three (conceptualization, design, and modeling) can be performed *in silico*. Designing and modeling synthetic biology applications *in silico* holds several advantages including speed, tractability, advantages associated with certain types of mathematics such as optimization, and the potential to develop a pipeline for synthetic biology applications. This has been recognized by other researchers, who have developed *in silico* tools for either design or modeling of genetic circuits, which are generally not paired with a complimentary tool in the other step (see Figure 5.1). This work seeks to address this lack, as well as expanding and improving upon optimization-based circuit design algorithms. In this work, it was decided to design and model plant-based heavy metal ion biosensors in Arabidopsis. These biosensors were designed to detect Cadmium, Copper, and Zinc, which are common metal ion pollutants, as a potential basis for future synthetic biology applications for phytoremediation. Arabidopsis was chosen as a model plant system with many previous synthetic biology applications, and it is eventually intended to apply EuGeneCiD and EuGeneCiM tools for applications in other plants (e.g., maize).

In the current work, two optimization-based tools for the design and modeling steps of the development of synthetic biology applications are introduced, the Eukaryotic Genetic Circuit Design (EuGeneCiD) and Modeling (EuGeneCiM) tools. The first tool uses inputs of a bioparts database and a conceptualization of the desired application (in the form of a logic table) to design genetic circuits. This tool is unique compared to previous tools in that it models transcript production; focuses on eukaryotic systems; accounts for transcript and enzyme degradation; and is more granular in its predictions than previous optimization-based tools. EuGeneCiD is paired with the dynamic circuit modeling tool EuGeneCiM, which uses the EuGeneCiD design and the bioparts databases as inputs. See Figure 5.4 for a visualization of the workflow.

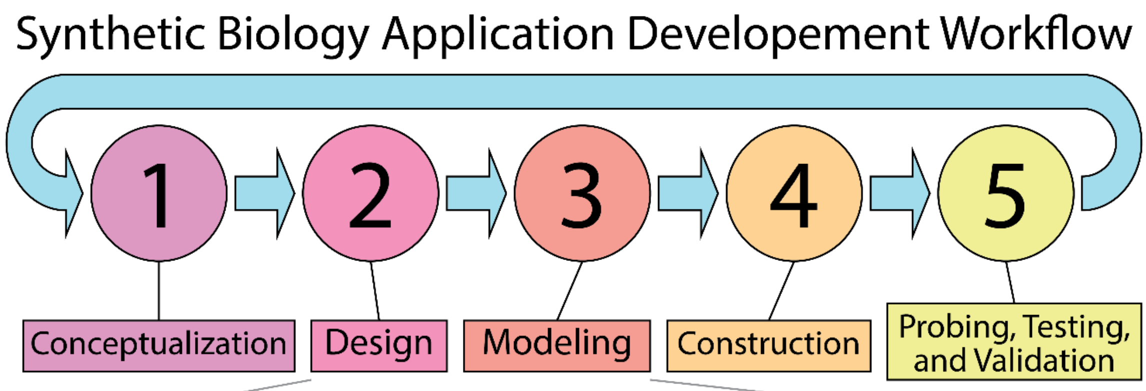
Once these tools were developed, they were applied to 27 different systems biology conceptualizations which were created by pairing a logic gate (AND, NIMPLY CNI, HALF ADDER, NAND, NOR, OR, XNOR, and XOR) with a pair of ligands for that gate to respond to (Cd/Cu, Cd/Zn, and Cu/Zn). These conceptualizations were chosen so as to make Arabidopsis roots as biosensors for heavy metal pollution, which can eventually be used as a basis for synthetic biology phytoremediation applications. EuGeneCiD and EuGeneCiM were run for seven days for each of the 27 conceptualization. The results of this are shown broadly in Figure 5.5, with some specific solutions to both EuGeneCiD and EuGeneCiM shown in Figure 5.6. Briefly, EuGeneCiM solves more quickly and with higher fractions of optimal solutions for simpler circuit logic, for example AND, NIMPLY, and NOR and more slowly for more difficult logics like XOR, HALF ADDER, and XNOR. As shown in Figure 5.6, when modeled dynamically, while many EuGeneCiD-created designs functioned correctly, designs did not always function correctly under dynamic modeling. This showed that EuGeneCiM adds value by screening potentially unsuccessful solutions. This is in part because EuGeneCiD does not design circuits with respect to time, so accumulation of enzymes and transcripts are not accounted for at the design stage. We also wished to emphasize that the EuGeneCiM tool could be used as a stand-alone dynamic genetic circuit modeling tool, and to this end, EuGeneCiM is successfully applied to a manually designed repressilator (see Figure 5.7). This highlights how the EuGeneCiM tool crucially accounts for enzyme and transcript degradation allowing modeling of important dynamic circuits such as repressilators.

As shown in Figure 5.7, no set of EuGeneCiD solutions for any of the 27 synthetic biology application conceptualizations produced only optimal solutions. For all, some fraction of solutions were integer solutions with no guarantee of optimality (local or global). The conceptualization with the highest fraction of optimal solutions is the Cd^{2+}/Zn^{2+} responsive AND circuit with 84% and that with the lowest fraction is the Cd^{2+}/Cu^{2+} responsive HALF ADDER and XOR circuits with slightly less more than 9% of solutions being optimal. The lack of any conceptualization identifying only optimal solutions has a few possible explanations. The first is that there is some “best” set of solver settings which would achieve only optimal solutions which we have not been able to identify. Due to the long run time of some circuit designs (seven days), it was not deemed worth the time and effort to identify this set. A second possibility is the sheer number of solutions sought in that the runs were set only to terminate when 1000 solutions had been identified, the sought circuit size exceeded ten triads, or seven days had passed. In the output of EuGeneCiD, it was found that for the Cd^{2+}/Zn^{2+} responsive AND circuit, of the 160 non-optimal solutions returned, 81 of these occur in the last 150 solutions identified. Other non-optimal solutions occur when only a single solution remains at a given circuit size. In some instances, a non-optimal solution code might also be returned for a solution with the same objective value as an immediately preceding optimal solution (to two decimal points), suggesting that in some cases the non-optimality is inconsequential. Similar patterns occur for many of the easy to solve conceptualizations such as AND, NIMPLY, and CNI. By this point, a large number of integer cuts have been defined in the model to prevent repeat solutions, increasing the difficulty of finding a solution. When more difficult, this result is longer run times and an increased likelihood of heuristic termination from the solver. These heuristic terminators include lack of improvement on solution bounds in a certain time frame and reaching the maximum allowed time for a single solution (set at 1E4 seconds). These heuristic terminations also might explain the differences between optimality ratios, such as between the Cd^{2+}/Zn^{2+} responsive AND and Cd^{2+}/Cu^{2+} responsive HALF ADDER circuits, in that solving the latter is significantly more difficult than the former. Given the relative positions of optimal to non-optimal solutions, the positions of solutions with the maximum objective value, and the lengthening solution times at higher solution numbers, for users of the EuGeneCiD tool it is recommended that only the first 100 solutions need be identified and investigated.

As noted earlier, EuGeneCiD is not a dynamic design tool, though it does attempt to model one half-lives degradation to attempt to overcome this issue and to include degradation in design criteria. This results in some design solutions being non-functional under dynamic modeling in EuGeneCiM. EuGeneCiD was made non-dynamic for one primary reason: computational expense. Given the number of binary variables inherent in the EuGeneCiD problem, the already long solution times for certain conceptualizations, and the frequent non-optimality of solutions, it was decided not to create a dynamic EuGeneCiD out of concern for creating a non-viable tool (or one viable only in niche instances). In future, it is desired to improve the EuGeneCiD tool, and one of the primary improvements we will aim to implement is to make the tool dynamic, potentially creating a hybrid design and modeling tool. Another issue arising from pairing a static and dynamic tool such as this, is the cumulative effects of concentration buildup in the dynamic model. This resulted in the need to halve terminator and enzyme half-lives to attempt to reach similar enzyme production levels in EuGeneCiD as in EuGeneCiM. Without this adjustment, EuGeneCiM predicted levels often were one to two order of magnitude larger than in EuGeneCiD, resulting in all enzymes in the design being "active" regardless of regulation. This approach to reduce the half-live seemed best to both minimize the changes the parameters (such as enzyme concentration level thresholds, half-life, transcriptional efficiency, etc.) and to still produce results on a similar order of magnitude.

Overall, EuGeneCiD and EuGeneCiM have the potential to design with respect to and model biopart interactions which do not exist in the current bioparts database. Some of these functionalities include alternative splicing, changeable transcriptional efficiency (such as might be tuned through codon optimization), and protein-protein regulatory interactions. In creating a more capable tool, we hope to encompass new bioparts with sophisticated functionality and regulation which are even now being created by synthetic biologists for fine-tuned control of designed systems. One example is the Two-Component Systems (TCSs) for phosphoregulated, chemically induced signal transduction in mammalian cells, a work which shows great potential for the future designs of sophisticated synthetic biology bioparts (Scheller et al., 2020). In addition to making EuGeneCiD and EuGeneCiM potentially compatible with future synthetic bioparts, the choice of system and knowledge of that system has limited the biopart interactions which might be present in the library. Arabidopsis was chosen as a test system because it is a model plant to which synthetic biology applications have previously been applied. A plant system was chosen for the application because, in future, we hope to use the EuGeneCiD and EuGeneCiM tools to create synthetic biology applications for *Zea mays*, particularly those which activate in response to stress conditions to increase plant health and fitness under these conditions. One potential application is for heavy metal phytoremediation, hence the use of heavy metal ligands as signals for designed genetic circuits. Given these desired goals and future applications, the breadth and types of interactions in the bioparts database was further limited.

5.5. FIGURES

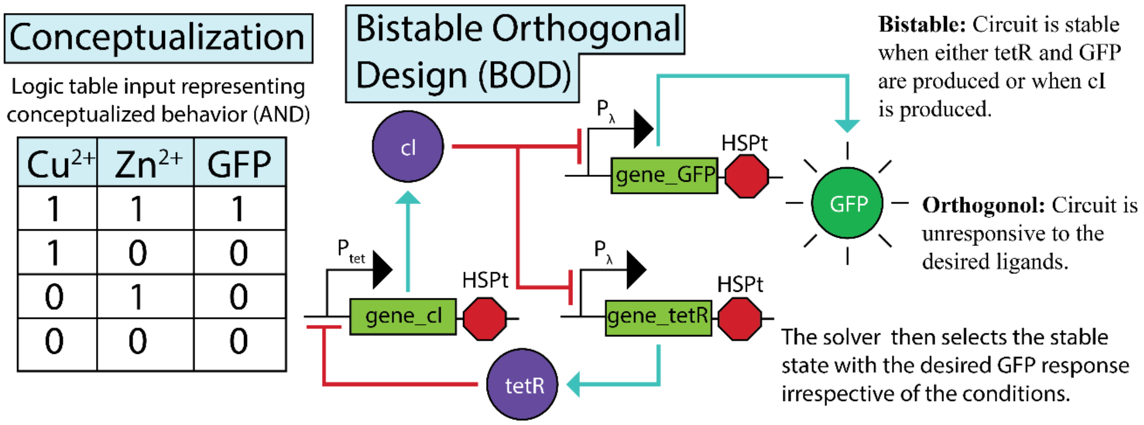

Synthetic Biology Application Developement Workflow
1
2
3
4
5
Conceptualization
Design
Modeling
Construction
Probing, Testing, and Validation

In Silico Tools

Design	Modeling
Cello (Chen et al., 2020) + • Can be highly accurate. • Aids plasmid design. — • Requires precise characterization. • Not "best" design.	No associated modeling tool. In Chen et al, 2020, this step is skipped, proceeds directly to step 5.
Zomorrodi and Maranas, 2014 + • Finds "best" design. • Uses relative values. — • Relative results. • Requires programming knowledge.	No associated modeling tool. In Zomorrodi and Maranas, 2014, the design stage was the last workflow step completed.
No associated design tool. In Davidsohn et al., 2015, the design was provided to the tool. The design process was not discussed.	EQuIP (Davidsohn et al., 2015) + • Incorporates copy number • Incorporates phenotype distribution. • Adaptable to many species. — • Requires precise characterization. • Utilizes difference equations.
EuGeneCiD (this work) + • Uses relative values. • Models transcript abundace. • Associated modelling tool. — • No plasmid design aid. • Relative results.	EuGeneCiM (this work) + • Associated design tool. • Uses a different objective. • Accounts for degradation. — • Ignores copy number. • Ignores phenotype distribution. • Relative results.

Figure 1: Steps of Synthetic Biology Application Development and some in silico Tools.

Extended Caption: Synthetic biology applications generally have five steps: conceptualization, design, modeling, construction, and probing, testing, and validation. Of these steps, three can be performed *in silico*. Several independent design and modeling tools exist for the second and third stages of this workflow, including Cello, the work of Zomorrodi and Maranas (2014) (in addition to their previous OptCircuit), and EQuIP. Introduced here are the EuGeneCiD and EuGeneCiM tools which integrate the design and modeling steps as design solutions are passed from EuGeneCiD to be modeled by EuGeneCiM. For the listed tools, a short list of strengths and weaknesses is included to help better position this work in the context of the current state of the field.



Cu^{2+}	Zn^{2+}	GFP
1	1	1
1	0	0
0	1	0
0	0	0

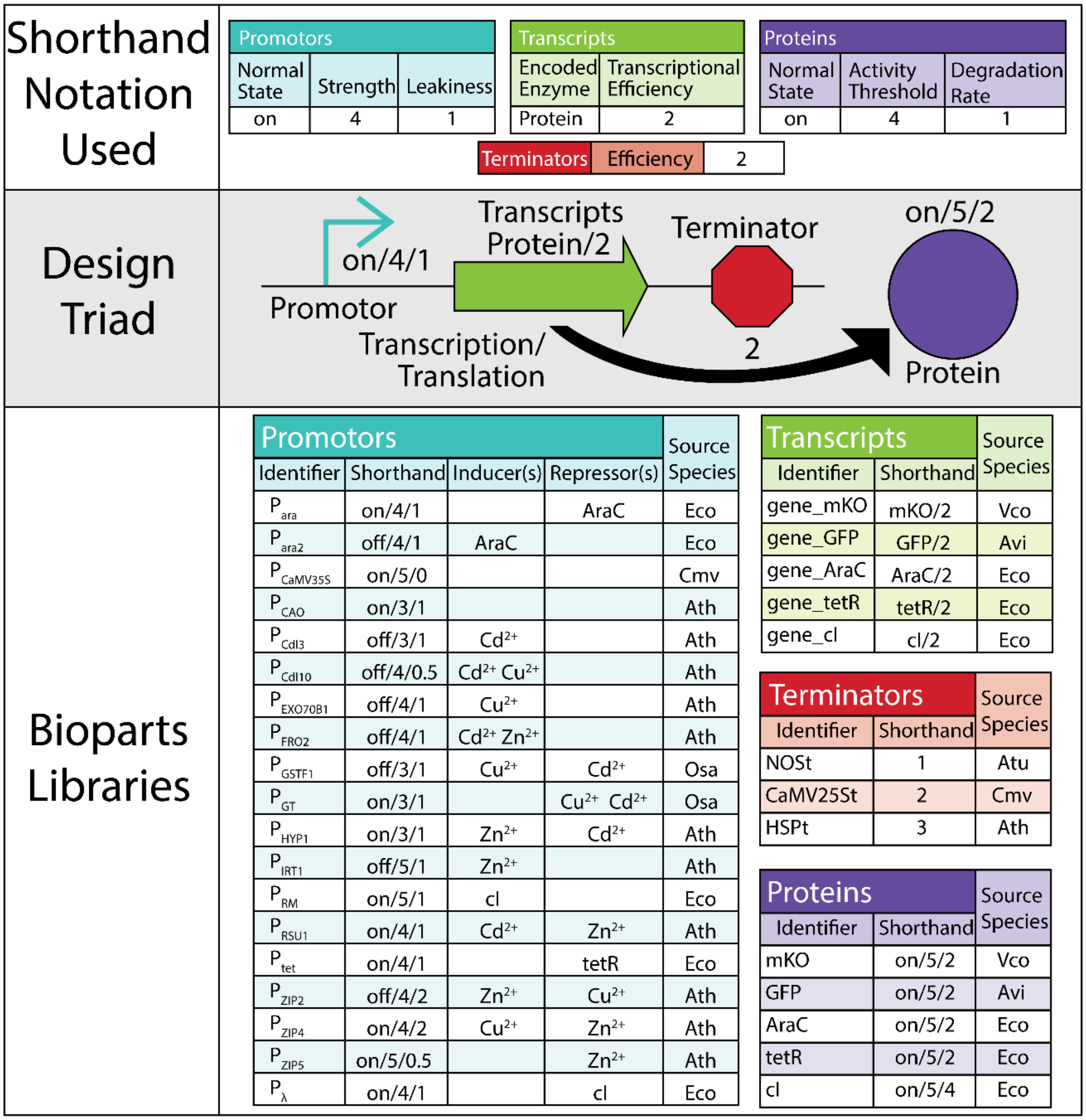
Bistable: Circuit is stable when either tetR and GFP are produced or when cI is produced.

Orthogonol: Circuit is unresponsive to the desired ligands.

The solver then selects the stable state with the desired GFP response irrespective of the conditions.

Figure 2: Example Bistable Orthogonal Design (BOD.

Extended Caption: This figure illustrates a major category of problematic potential designs which may be produced by optimization-based genetic circuit design tools. From a conceptualized Cu^{2+}/Zn^{2+} responsive AND circuit, it is possible, without attribution equations, to create Bistable Orthoganol Design (BOD) which can produce the desired response, yet not be responding to the desired signals. Text in the image describes why this occurs. One of the major innovations in EuGeneCiD is the development of attribution equations to avoid BODs.



Shorthand Notation Used

Promotors		
Normal State	Strength	Leakiness
on	4	1

Transcripts	
Encoded Enzyme	Transcriptional Efficiency
Protein	2

Proteins		
Normal State	Activity Threshold	Degradation Rate
on	4	1

Terminators	Efficiency	2

Design Triad

on/4/1
Promotor
Transcripts
Protein/2
Transcription/
Translation
Terminator
2
on/5/2
Protein

Bioparts Libraries

Promotors				Source
Identifier	Shorthand	Inducer(s)	Repressor(s)	Species
P_{ara}	on/4/1		AraC	Eco
P_{ara2}	off/4/1	AraC		Eco
$P_{CaMV35S}$	on/5/0			Cmv
P_{CAO}	on/3/1			Ath
P_{CdI3}	off/3/1	Cd^{2+}		Ath
P_{CdI10}	off/4/0.5	Cd^{2+} Cu^{2+}		Ath
$P_{EXO70B1}$	off/4/1	Cu^{2+}		Ath
P_{FRO2}	off/4/1	Cd^{2+} Zn^{2+}		Ath
P_{GSTF1}	off/3/1	Cu^{2+}	Cd^{2+}	Osa
P_{GT}	on/3/1		Cu^{2+} Cd^{2+}	Osa
P_{HYP1}	on/3/1	Zn^{2+}	Cd^{2+}	Ath
P_{IRT1}	off/5/1	Zn^{2+}		Ath
P_{RM}	on/5/1	cI		Eco
P_{RSU1}	on/4/1	Cd^{2+}	Zn^{2+}	Ath
P_{tet}	on/4/1		tetR	Eco
P_{ZIP2}	off/4/2	Zn^{2+}	Cu^{2+}	Ath
P_{ZIP4}	on/4/2	Cu^{2+}	Zn^{2+}	Ath
P_{ZIP5}	on/5/0.5		Zn^{2+}	Ath
P_{λ}	on/4/1		cI	Eco

Transcripts		Source
Identifier	Shorthand	Species
gene_mKO	mKO/2	Vco
gene_GFP	GFP/2	Avi
gene_AraC	AraC/2	Eco
gene_tetR	tetR/2	Eco
gene_cI	cI/2	Eco

Terminators		Source
Identifier	Shorthand	Species
NOSt	1	Atu
CaMV25St	2	Cmv
HSPt	3	Ath

Proteins		Source
Identifier	Shorthand	Species
mKO	on/5/2	Vco
GFP	on/5/2	Avi
AraC	on/5/2	Eco
tetR	on/5/2	Eco
cI	on/5/4	Eco

Figure 3: Bioparts Database for the Current Work.

Extended Caption: The EuGeneCiD and EuGeneCiM Tools designed require the definition of bioparts databases from which to pick design elements and to define the properties of those elements for both design and modeling. For compactness in other images, introduced here is a shorthand for promotor, transcript, terminator, and protein characteristics. The shorthand here is then used to define each biopart included in the bioparts library used for this work, which includes promotors, transcripts, terminators, and proteins. Source species acronyms for listed bioparts are as follows: Ath – *Arabidopsis thaliana*, Osa – *Oryzae sativa*, Eco – *Escherichia coli*, Vco – *Verrillofungia coninna*, Avi – *Aequorea victoria*, Atu – *Agrovacterium tumefaciens*, Cmv – Califlower Mosaic Virus.

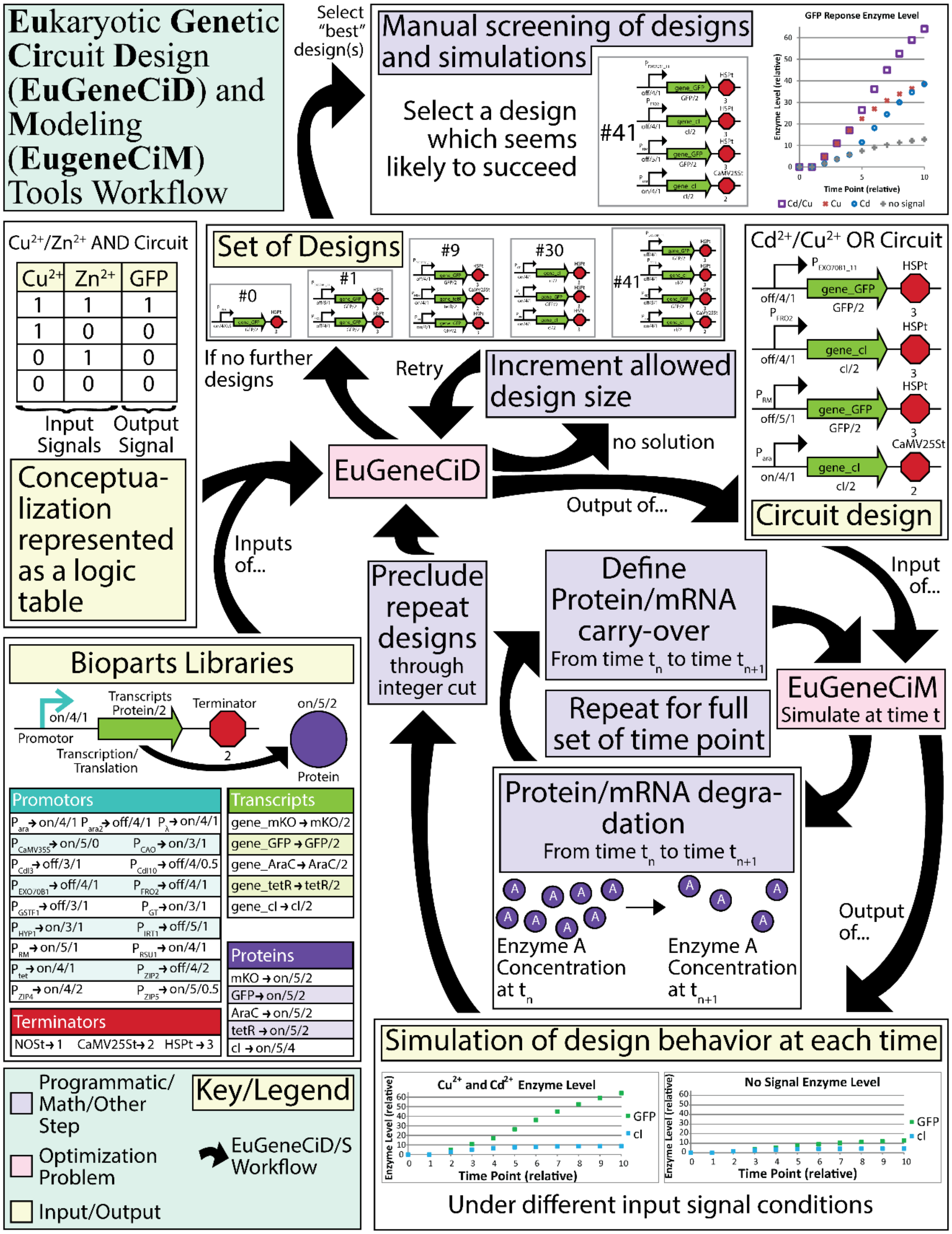
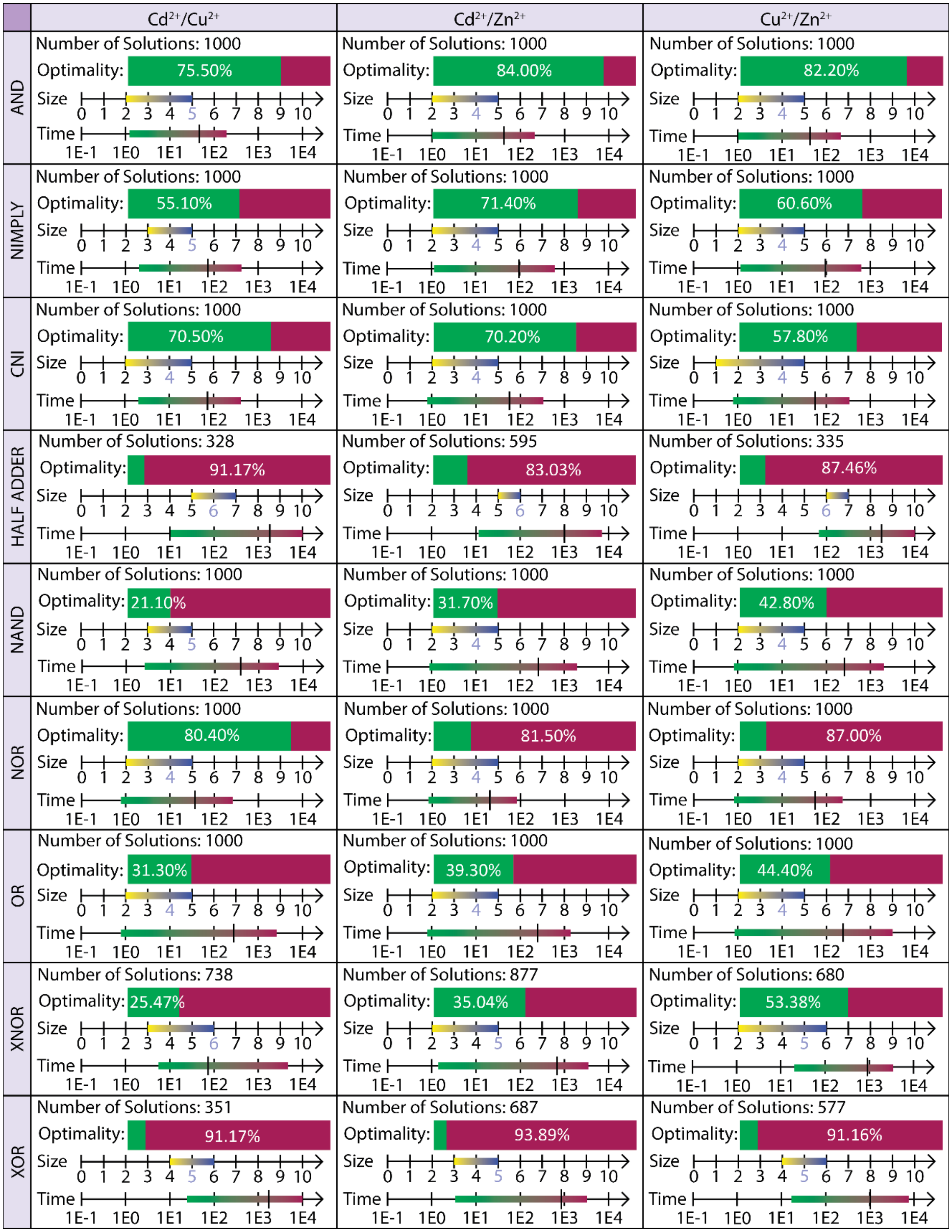

Eukaryotic Genetic Circuit Design (EuGeneCiD) and Modeling (EugeneCiM) Tools Workflow
Select "best" design(s)
Manual screening of designs and simulations
Select a design which seems likely to succeed
#41
GFP Reponse Enzyme Level
Enzyme Level (relative)
Time Point (relative)
Cd/Cu
Cu
Cd
no signal
Cu2+/Zn2+ AND Circuit
Cu2+
Zn2+
GFP
1 1 1
1 0 0
0 1 0
0 0 0
Input Signals
Output Signal
Conceptualization represented as a logic table
Set of Designs
#0
#1
#9
#30
#41
If no further designs
Retry
Increment allowed design size
no solution
EuGeneCiD
Output of...
Inputs of...
Cd2+/Cu2+ OR Circuit
P EXO70B1_11
P FRO2
P RM
P ara
off/4/1
off/4/1
off/5/1
on/4/1
gene_GFP
gene_cl
GFP/2
cl/2
HSPt
3
CaMV25St
2
Circuit design
Preclude repeat designs through integer cut
Define Protein/mRNA carry-over From time t_n to time t_n+1
Input of...
EuGeneCiM Simulate at time t
Repeat for full set of time point
Protein/mRNA degradation From time t_n to time t_n+1
Enzyme A Concentration at t_n
Enzyme A Concentration at t_n+1
Output of...
Bioparts Libraries
on/4/1
Transcripts Protein/2
Terminator
on/5/2
Promotor
Transcription/ Translation
2
Protein
Promotors
P_ara → on/4/1 P_ara2 → off/4/1 P_λ → on/4/1
P_CaMV35S → on/5/0 P_CAO → on/3/1
P_CdI3 → off/3/1 P_CdI10 → off/4/0.5
P_EXO70B1 → off/4/1 P_FRO2 → off/4/1
P_GSTF1 → off/3/1 P_GT → on/3/1
P_HYP1 → on/3/1 P_IRT1 → off/5/1
P_RM → on/5/1 P_RSU1 → on/4/1
P_tet → on/4/1 P_ZIP2 → off/4/2
P_ZIP4 → on/4/2 P_ZIP5 → on/5/0.5
Terminators
NOSt → 1 CaMV25St → 2 HSPt → 3
Transcripts
gene_mKO → mKO/2
gene_GFP → GFP/2
gene_AraC → AraC/2
gene_tetR → tetR/2
gene_cl → cl/2
Proteins
mKO → on/5/2
GFP → on/5/2
AraC → on/5/2
tetR → on/5/2
cl → on/5/4
Simulation of design behavior at each time
Cu2+ and Cd2+ Enzyme Level
No Signal Enzyme Level
Enzyme Level (relative)
Time Point (relative)
GFP
cl
Under different input signal conditions
Key/Legend
Programmatic/ Math/Other Step
Optimization Problem
Input/Output
EuGeneCiD/S Workflow

Figure 4: Workflow of the EuGeneCiD and EuGeneCiM Tools.

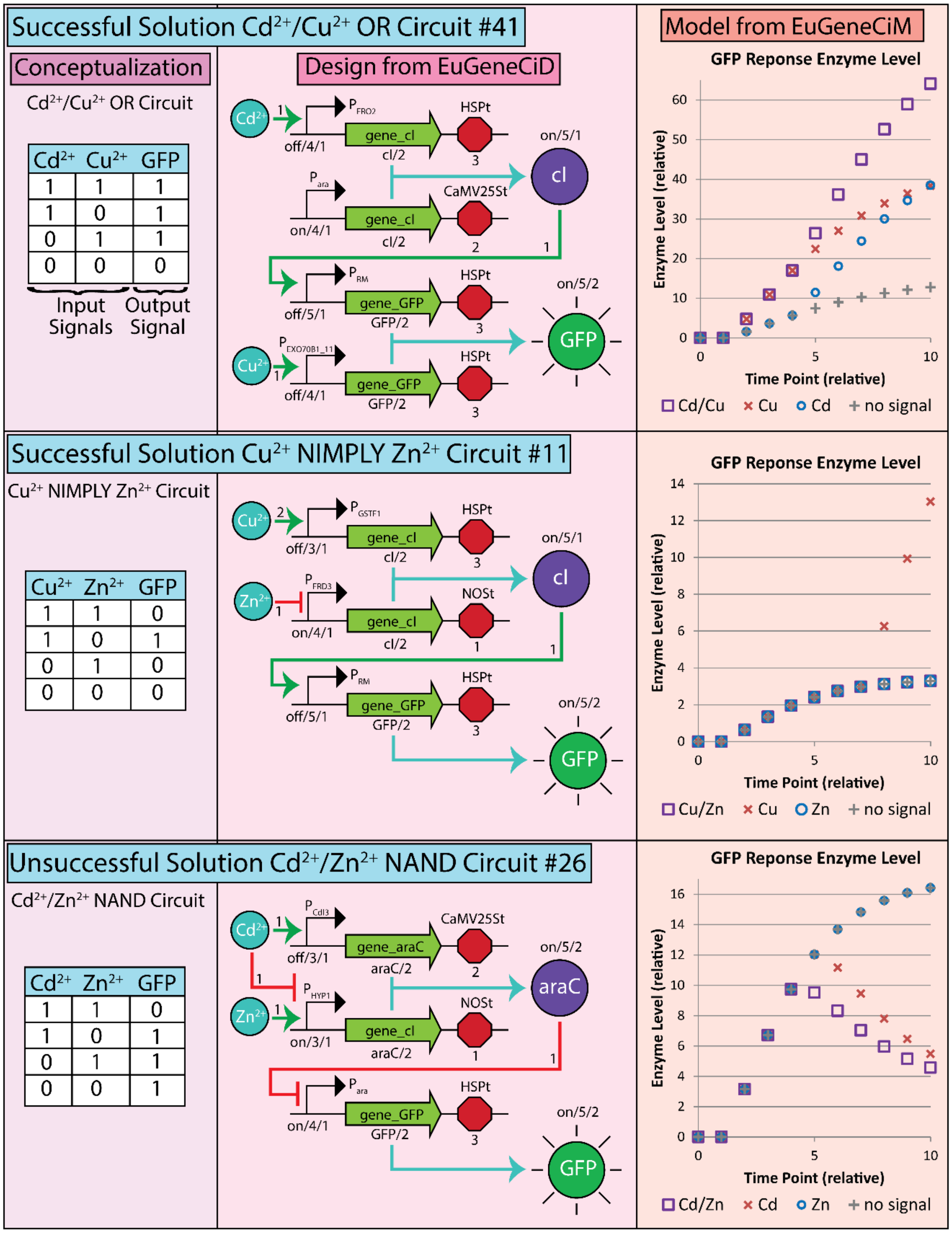
Extended Caption: The EuGeneCiD and EuGeneCiM tools were designed to be used in concert to complete the design and modeling steps of synthetic biology applications development together. This workflow begins with a defined conceptualization of the application (in the form of a logic table) and a bioparts library which defines and describes potential design elements (see Figure 2). Then an attempt to solve EuGeneCiD is made, with three possible outcomes. First, no solution is found at the current design size limit (limiting the number of allowed triads), in which case this limit is incremented, and EuGeneCiD is attempted again. Should design or run limits be reached, or if no further designs exist within specified restrictions, the set of designs is returned which can be manually screened for candidates likely to succeed. Should the attempt to solve EuGeneCiD be successful, a circuit design is the result, which is passed to EuGeneCiM for modeling. This modeling solves EuGeneCiM at each time point and applies protein and transcript degradation between time points for the full set of desired model time points. This results in a simulation of design behavior at each time point which will be reported. The current solution is then precluded by defining a new integer cut and the cycle is repeated.



	Cd^{2+}/Cu^{2+}	Cd^{2+}/Zn^{2+}	Cu^{2+}/Zn^{2+}
AND	Number of Solutions: 1000 Optimality: 75.50%	Number of Solutions: 1000 Optimality: 84.00%	Number of Solutions: 1000 Optimality: 82.20%
NIMPLY	Number of Solutions: 1000 Optimality: 55.10%	Number of Solutions: 1000 Optimality: 71.40%	Number of Solutions: 1000 Optimality: 60.60%
CNI	Number of Solutions: 1000 Optimality: 70.50%	Number of Solutions: 1000 Optimality: 70.20%	Number of Solutions: 1000 Optimality: 57.80%
HALF ADDER	Number of Solutions: 328 Optimality: 91.17%	Number of Solutions: 595 Optimality: 83.03%	Number of Solutions: 335 Optimality: 87.46%
NAND	Number of Solutions: 1000 Optimality: 21.10%	Number of Solutions: 1000 Optimality: 31.70%	Number of Solutions: 1000 Optimality: 42.80%
NOR	Number of Solutions: 1000 Optimality: 80.40%	Number of Solutions: 1000 Optimality: 81.50%	Number of Solutions: 1000 Optimality: 87.00%
OR	Number of Solutions: 1000 Optimality: 31.30%	Number of Solutions: 1000 Optimality: 39.30%	Number of Solutions: 1000 Optimality: 44.40%
XNOR	Number of Solutions: 738 Optimality: 25.47%	Number of Solutions: 877 Optimality: 35.04%	Number of Solutions: 680 Optimality: 53.38%
XOR	Number of Solutions: 351 Optimality: 91.17%	Number of Solutions: 687 Optimality: 93.89%	Number of Solutions: 577 Optimality: 91.16%

Figure 5: Visualized EuGeneCiD Results

This three by nine grid reports on the general characteristics of the set of EuGeneCiD results for each circuit conceptualization. From top to bottom of each grid, four items describing the results set are shown. First, is the number of solutions in that set. Second, is the percentage of results which are optimal (if this value is above 20%, green bar) or the percentage of results that are suboptimal (red, if the value of this is above 80%). Third is a number line, which indicates the solution set minimum and maximum sizes (in the number of triads in the design) and the mode size (the number is shaded blue). This number line is extended from zero to ten as ten is the maximum allowed circuit size (though no solution was created of this size). Finally, another number line shows the minimum and maximum solution times (in seconds) on a logarithmic scale. A large black line on the solution time range indicates the mean solution time.



Cd^{2+}	Cu^{2+}	GFP
1	1	1
1	0	1
0	1	1
0	0	0

Cu^{2+}	Zn^{2+}	GFP
1	1	0
1	0	1
0	1	0
0	0	0

Cd^{2+}	Zn^{2+}	GFP
1	1	0
1	0	1
0	1	1
0	0	1

Figure 6: Example EuGeneCiD and EuGeneCiM Solutions.

Extended Caption: Shown here are three circuit conceptualizations, EuGeneCiD design solutions, and their associated EuGeneCiM models. The conceptualization is shown as the input logic table. The solution is shown with the design triads and produced enzymes with regulatory relations shown (green for activation, red for inhibition), including their relative strengths (shown as numbers on top of the regulation line). The modeled design responses are shown in the rightmost panel; where purple squares indicate the presence of both signals; blue circles and red crosses denote only one signal (see individual legends); and grey plus signs indicate no signal. Of the provided solutions, two of were shown to be potentially successful (Cd^{2+}/Cu^{2+} OR circuit solution #41 and Cu^{2+} NIMPLY Zn^{2+} circuit solution #11) and one shown to be potentially unsuccessful (Cd^{2+}/Zn^{2+} NAND Circuit solution #26) by EuGeneCiM.

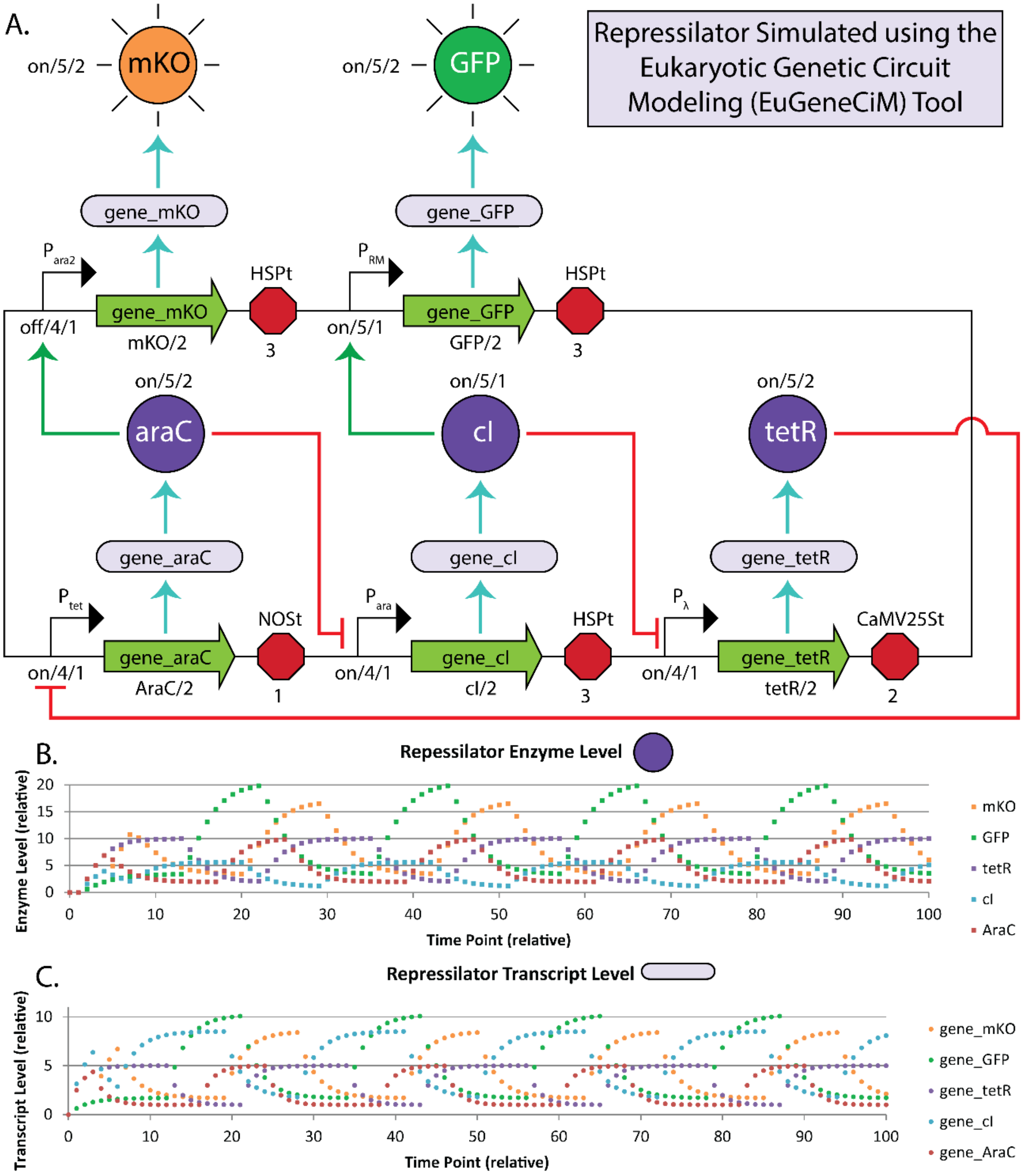


Figure 7: Repressilator simulated using the EuGeneCiM Tool.

Extended Caption: While the EuGeneCiD and EuGeneCiM tools were designed to use in concert, they can be used independently, as evidenced here where EuGeneCiM is used to model a manually-designed repressilator. A) Shows the repressilator design with promotors (black), transcripts (green), and terminators (red) (collectively the design triads) in addition to the transcripts (light purple) and proteins (purple) produced thereby. The shorthand used throughout this work is used to show the characterization of these parts. Further, regulatory relations are shown (green for activation, red for inhibition). B) Scatter plot showing the dynamic behavior of the enzyme level for each of the enzymes included in the repressilator. C) Scatter plot showing the dynamic behavior of the transcript level for each of the enzymes included in the repressilator.

5.6. METHOD DETAILS

5.6.1. Symbols Used

This section is provided here to increase clarity of the provided equations which follow. For the purposes of this text, a set is an unordered collection of distinct elements, a parameter is a value which is constant during the solution process whereas the value of a variable is altered by the solver to identify optimal solutions.

5.6.1.1. Sets

$A \equiv set\ of\ all\ molecules$

$P \subset A \equiv set\ of\ promotors$

$J \subset A \equiv set\ of\ transcripts$

$E \subset A \equiv set\ of\ enzymes$

$E_d \subseteq E \equiv set\ of\ enzymes\ which\ it\ is\ desired\ for\ the\ circuit\ to\ respond\ to$

$L \subset A \equiv set\ of\ ligands$

$L_d \subseteq L$

$\equiv set\ of\ ligands\ which\ it\ is\ desired\ for\ the\ circuit\ to\ respond\ to\ (note\ that\ this\ should\ always\ contain\ a$ none)

$T \subset A \equiv set\ of\ all\ terminators$

$\mathbb{R} \equiv set\ of\ real\ numbers$

$\mathbb{R}^+ \equiv set\ of\ nonnegative, real\ numbers$

$\mathbb{R}^{-} \equiv set\ of\ nonpositive, real\ numbers$

$\mathbb{B} \equiv binary\ set, contains\ only\ the\ numbers\ 1\ and\ 0, e.g. \mathbb{B} = \{0,1\}$

$\mathbb{T} \equiv trinary\ set\ containing\ only\ the\ numbers - 1, 0, and\ 1, e.g. \mathbb{T} = \{-1,0,1\}$

5.6.1.2. Parameters

$\lambda_{eL_1L_2} \in \mathbb{B} \equiv input\ logic\ matrix\ value\ for\ enzyme\ e\ under\ conditions\ of$

$ligands\ L_1, L_2 \in L_d\ present$

$Z_p \in \mathbb{B} \equiv normal\ state\ of\ promotor\ p \in P$

$\zeta_e \in \mathbb{B} \equiv normal\ state\ of\ enzyme\ e \in E$

$I_{pa} \in \mathbb{T} \equiv Effects\ of\ a \in A\ as\ a\ ligand\ upon\ the\ activity\ of\ promotor\ p \in P$

$(-1\ inhibition, 0\ no\ effect, 1\ activation)$

$H_{pa} \in \mathbb{R}^{+} \equiv strength\ of\ interaction\ between\ promotor\ p \in P\ and\ molecule\ a$

$\in A$

$B_{ea} \in \mathbb{T} \equiv Effects\ of\ a \in A\ as\ a\ ligand\ upon\ the\ activity\ of\ enzyme\ e \in E$

$(-1\ inhibition, 0\ no\ effect, 1\ activation)$

$Q_{ea} \in \mathbb{R}^{+} \equiv strength\ of\ interaction\ between\ enzyme\ e \in E\ and\ molecule\ a \in A$

$V = 1E4 \equiv an\ arbitrarily\ large\ number$

$\epsilon = 1E - 4 \equiv an\ arbitrarily\ small\ number$

$\theta_e \in \mathbb{R}^{+} \equiv concentration\ threshold\ at\ which\ the\ enzyme\ e \in E\ must\ be\ present$

$to\ be\ said\ to\ be\ "active"$

$\eta_j \in \mathbb{R}^{+} \equiv translational\ efficiency\ of\ transcript\ j \in J$

$F_p \in \mathbb{R}^{+} \equiv leakiness\ of\ a\ promotor\ p \in P$

$G_t \in \mathbb{R}^+ \equiv half - life\ of\ terminator\ t \in T$

$\tau_e \in \mathbb{R}^+ \equiv half - life\ of\ enzyme\ e \in E$

$\sigma_{a_1 a_2} \in \mathbb{B} \equiv value\ of\ 1\ if\ a_1 \in A\ is\ the\ same\ as\ a_2 \in A\ and\ zero\ otherwise, identifies\ equivalent\ elements$

$S_p \in \mathbb{R}^+ \equiv strength\ of\ promotor\ p \in P$

5.6.1.3. Variables

$\alpha_{pL_1L_2} \in \mathbb{R} \equiv integer\ net\ effect\ of\ all\ inhibition\ and\ activation\ on\ a\ given\ promotor\ p \in P\ under\ conditions\ of\ ligands\ L_1, L_2 \in L_d\ present\ (> 0\ promotor\ can\ be\ active, \leq 0\ promotor\ cannot\ be\ active)$

$\alpha^+_{pL_1L_2} \in \mathbb{B} \equiv binary\ net\ effect\ of\ ligands\ upon\ promotor\ p \in P\ in\ circuit\ under\ ligand\ conditions\ L_1\ , L_2 \in L_d\ (1\ promotor\ can\ be\ active, 0\ promotor\ cannot\ be\ active)$

$\gamma_{eL_1L_2} \in \mathbb{R} \equiv integer\ net\ effect\ of\ all\ inhibition\ and\ ativation\ on\ a\ given\ enzyme\ e \in E\ under\ ligand\ conditions\ L_1, L_2 \in L_d\ (> 0\ enzyme\ can\ be\ active, \leq 0\ enzyme\ cannot)$

$\gamma^+_{eL_1L_2} \in \mathbb{B} \equiv binary\ net\ effect\ of\ ligands\ upon\ enyme\ e \in E\ in\ circuit\ under\ ligand\ conditions\ L_1, L_2 \in L_d\ (1\ enzyme\ can\ be\ active, 0\ enzyme\ cannot\ be\ active)$

$\phi_{jtL_1L_2} \in \mathbb{R}^+ \equiv level\ of\ transcript\ j\ expression\ under\ ligand\ conditions\ L_1, L_2 \in L_d$

$M_{pjt} \in \mathbb{B} \equiv$ *binary variable which creates promotor* $p \in P$, *transcript* $j \in J$, *and terminator* $t \in T$ *triads representing the design* (*variable in EuGeneCiD*, (*parameter in EuGeneCiS*)

$C_{eL_1L_2} \in \mathbb{R}^+ \equiv$ *concentration of enzyme* $e \in E$ *under under ligand conditions* $L_1, L_2 \in L_d$

$\xi_{pjtL_1L_2} \in \mathbb{R}^+ \equiv$ *deliberate transcription of* $j \in J$ *transcribed from promotor* $p \in P$ *and transcript* $t \in T$ *under ligand conditions* $L_1, L_2 \in L_d$

$\omega_{eL_1L_2} \in \mathbb{B} \equiv$ *determines if enzyme* $e \in E$ *is produced under ligand conditions* $L_1, L_2 \in L_d$

$Y_{eL_1L_2} \in \mathbb{B} \equiv$ *binary variable determining if the enzyme* $e \in E$ *has sufficient concentration to be considered* active *under ligand conditions* $L_1, L_2 \in L_d$

$W_{eL_1L_2} \in \mathbb{B} \equiv$ *binary variable determining if enzyme* $e \in E$ *is both at sufficient concentration to be active and that it is not inhibited, in short that it will function underligand conditions* $L_1, L_2 \in L_d$

$\kappa_{eL_1L_2} \in \mathbb{B} \equiv$ *binary variable determining if enzyme* $\in E$ *is produced and can be active under ligand conditions* $L_1, L_2 \in L_d$

$Z_D \in \mathbb{R} \equiv$ *objective variable for EuGeneCiD*

$Z_M \in \mathbb{R} \equiv$ *objective variable for EuGeneCiM*

$D_{ee_1} \in \mathbb{R}^+ \equiv$ *direct attribution of enzyme* e *activity to* e_1 *through enzyme interactions*

$K_{ee_1} \in \mathbb{R}^+ \equiv$ *direct attribution of enzyme* e *activity to* e_1 *through enzyme* e_1 *on triad interactions*

$U_{ee_1e_2} \in \mathbb{R}^+ \equiv attribution\ of\ enzyme\ e\ activity\ to\ e_2\ acting\ through\ e_1\ through\ enzyme\ interactions$

$U'_{ee_1} \in \mathbb{R}^+ \equiv networked\ attribution\ of\ enzyme\ e\ activity\ to\ e_1\ acting\ through\ other\ enzymes\ reflecting\ direct\ enzyme - enzyme\ interactions$

$\chi_{ee_1e_2} \in \mathbb{R}^+ \equiv attribution\ of\ enzyme\ e\ activity\ to\ e_2\ acting\ through\ e_1\ through\ enzyme\ on\ triad\ interactions$

$X_{ee_1} \in \mathbb{R}^+ \equiv networked\ attribution\ of\ enzyme\ e\ activity\ to\ e_1\ acting\ through\ other\ enzymes\ reflecting\ enzyme\ on\ triad\ interactions$

$L_e \in \mathbb{B} \equiv value\ of\ 1\ if\ enzyme\ e\ is\ encoded\ by\ the\ genetic\ circuit\ design, 0\ otherwise$

$\beta_{ee_1} \in \mathbb{B} \equiv value\ of\ 1\ if\ enzymes\ e\ and\ e_1\ are\ encoded\ by\ the\ genetic\ circuit\ design, 0\ otherwise$

5.6.2. EuGeneCiD Problem Statement and Explanation

5.6.2.1. Objective function

5.6.2.1.1. Objective Function (equation 1)

$$\boldsymbol{maximize\ Z_D = \sum_{e \in E_d} \sum_{L_1 \in L_d} \sum_{L_2 \in L_d} \left[C_{eL_1L_2} \lambda_{eL_1L_2} - C_{eL_1L_2}\left(1 - \lambda_{eL_1L_2}\right)\right]} \tag{5.1}$$

Where Z_D is the objective value, $C_{eL_1L_2}$ is the contraction of enzyme e under conditions with signals L_1 and L_2 (which includes "none") and $\lambda_{eL_1L_2}$ is the desired phenotype in response to signals L_1 and L_2 as encoded in the conceptualized logic table (this term is order-dependent). See the methods section for the full list of symbols and their definitions. This equation, equation (5.1), seeks to maximize the responses of the desired enzymes under their desired conditions (in terms of concentration) and minimize the responses of the undesired enzymes under their undesired condition.

Note that in the above, the order of set elements matters, e.g. $C_{GFP,Zn^{2+},none}$ is mathematically distict from $C_{GFP,none,Zn^{2+}}$ though efforts have been made to ensure that they will have the same value. Nonetheless, the issue of combinations (of which there are a total of 8 for any given ligand set in this work, where the set includes the two ligands to which the system should respond as well as "none") affects the objective function. From this, an AND circuit would only have 1 of 8 values of $\lambda_{eL_1L_2}$ with a 1 and the remainder would be 0. Similarly, a NOR circuit would only have a single non-zero value in its order-dependent conceptualization matrix ($\lambda_{eL_1L_2}$). This results in these circuits having unusually low objective values, as most terms are subtractive. The tendency in optimal designs then is to strongly favor designs with minimal expression leakage. Conversely, OR and NAND circuits have only one or two zero values in their order-dependent conceptualization matrix ($\lambda_{eL_1L_2}$), and therefore most terms are additive. Therefore, optimal circuit designs here tend to favor high inducible expression. Therefore, in Figure 5.7, it is best to not compare objective function values between different conceptualizations, but to only compare within conceptualizations. Depending on the tendencies of circuit design due to the circuit type, more complex circuits could result in lower expression leakage or higher inducible expression, and these complexities cannot be built into small circuits consisting of one or two triads.

5.6.2.2. Constraint Equations

5.6.2.2.1. Circuit size limitations (equations 5.2 through 5.5)

These equations limit the number of:

1) Maximum number of copies of a single promotor which can be used in the circuit design ($N_{p,max}$), equation (5.2).
2) Maximum number of copies of a single transcript which can be used in the circuit design ($N_{j,max}$), equation (5.3).
3) Maximum number of copies of a single terminator which can be used in the circuit design ($N_{t,max}$), equation (5.4).
4) Total number of promotors, transcripts, and terminator triads which the circuit design can use ($N_{circuit,max}$), equation (5.5).

$$\sum_{j\in J}\sum_{t\in T}\boldsymbol{M_{pjt}} \leq \boldsymbol{N_{p,max}} \qquad \forall\, \boldsymbol{p} \in \boldsymbol{P} \qquad (5.2)$$

$$\sum_{p\in P}\sum_{t\in T}\boldsymbol{M_{pjt}} \leq \boldsymbol{N_{j,max}} \qquad \forall\, \boldsymbol{j} \in \boldsymbol{J} \qquad (5.3)$$

$$\sum_{p \in P} \sum_{j \in J} \boldsymbol{M_{pjt}} \leq \boldsymbol{N_{t,max}} \qquad \forall\, \boldsymbol{t} \in \boldsymbol{T} \tag{5.4}$$

$$\sum_{j \in J} \sum_{p \in P} \sum_{t \in T} \boldsymbol{M_{pjt}} \leq \boldsymbol{N_{circuit,max}} \tag{5.5}$$

Note that by the nature of the variables used (e.g., M_{pjt} being binary), only one copy of any given triad may be present in the designed circuit. However, any number of promotor/transcript, promotor/terminator, and transcript/terminator pairs may be repeated. This is important to later constraints. It should be noted that $N_{circuit,max}$ is set to 1 in the first attempt to solve EuGeneCiD and incremented by 1 each time no solution is found or the problem is deemed infeasible. In this way, the simplest circuit designs possible are identified and precluded from future solutions so that each solution is the simplest possible.

5.6.2.2.2. Promotor state under conditions (equations 5.6 through 5.8)

These equations determine if a promotor is active under the given conditions of ligand 1 and/or/nor 2 being present. Equations perform as follows:

1) Determines the net effect of (by term): i) promotor normal state, ii) activation or inhibition by enzymes produced by the circuit, iii) inhibition or activation by ligand L_1, iv) inhibition or activation by ligand L_2, v) prevent duplicate activation/inhibition if L_1 and L_2. Equation (5.6).
2) Ensures that if $\alpha_{pL_dL_{d1}} > 0$ then $\alpha^{+}_{pL_1L_2} = 1$, and if $\alpha_{pL_dL_{d1}} \leq 0$ then $\alpha^{+}_{pL_1L_2} = 0$. Equation (5.7) and (5.8).

$$\boldsymbol{\alpha_{pL_1L_2} = Z_p + \sum_{e \in E} \left[W_{eL_1L_2} I_{pe} H_{pe}\right] + I_{pL_1} H_{pL_1} + I_{pL_2} H_{pL_2} - I_{pL_1} \sigma_{L_1L_2} H_{pL_1}} \qquad \forall\, \boldsymbol{p \in P; L_1, L_2 \in L_d} \qquad (5.6)$$

$$\boldsymbol{\alpha_{pL_1L_2} \geq -V\left(1 - \alpha^{+}_{pL_1L_2}\right) + \epsilon \alpha^{+}_{pL_1L_2}} \qquad \forall\, \boldsymbol{p \in P; L_1, L_2 \in L_d} \qquad (5.7)$$

$$\boldsymbol{\alpha_{pL_1L_2} \leq V \alpha^{+}_{pL_1L_2}} \qquad \forall\, \boldsymbol{p \in P; L_1, L_2 \in L_d} \qquad (5.8)$$

5.6.2.2.3. Transcription under conditions (equations 5.9 through 5.12)

These equations determine if and to what extent transcript j is intentionally transcribed from promotor p under ligand L_1 and L_2 conditions ($\xi_{pjtL_1L_2}$). The following equations accomplish the following:

1) A transcript cannot be transcribed from a given promotor unless the promotor and transcript are paired in the circuit design.
2) Transcription won't occur unless the promotor is "on".
3) All three constraints are equivalent to: $\xi_{pjtL_dL_{d1}} = S_p M_{pjt} \alpha^{+}_{pL_1L_2}$, equations (5.9), (5.10), and (5.11).

$$\boldsymbol{\xi_{pjtL_1L_2} \leq S_p M_{pjt}} \qquad \forall\, \boldsymbol{p \in P; j \in J; t \in T; L_1, L_2 \in L_d} \qquad (5.9)$$

$$\boldsymbol{\xi_{pjtL_1L_2} \leq S_p \alpha^{+}_{pL_1L_2}} \qquad \forall\, \boldsymbol{p \in P; j \in J; t \in T; L_1, L_2 \in L_d} \qquad (5.10)$$

$$\boldsymbol{\xi_{pjtL_1L_2} \geq S_p\left(M_{pjt} + \alpha^{+}_{pL_1L_2} - 1\right)} \qquad \forall\, \boldsymbol{p \in P; j \in J; t \in T; L_1, L_2 \in L_d} \qquad (5.11)$$

The following equations determine the transcript level ($\phi_{jL_dL_{d1}}$) as the sum of positive effects on the transcript level, including deliberate ($\xi_{pjL_1L_2}$) and leaky ($M_{pjt}F_p$) transcription. This is scaled by a half-life-based amount of RNA degradation to simulate the fact that degradation occurs and factors this into circuit design.

$$\boldsymbol{\phi_{jtL_1L_2} = \sum_{p \in P} \left[\left(\xi_{pjL_1L_2} + M_{pjt}F_p \right) \left(0.5^{\left(\frac{1}{G_t + \epsilon}\right)} \right) \right] \qquad \forall\, j \in J; t \in T; L_1, L_2 \in L_d} \qquad (5.12)$$

5.6.2.2.4. Translation under conditions (equations 5.13 through 5.17)

The following equation determines the enzyme concentration level ($C_{eL_dL_{d1}}$) as the sum of effects on the enzyme concentration level ($C_{eL_dL_{d1}}$), equation (5.17), reduced by a half-life-based enzyme degradation multiplicative factor.

$$\boldsymbol{C_{eL_1L_2} = \sum_{j \in J} \left[\left(\rho_{je} \eta_j \phi_{jL_1L_2} \right) \left(0.5^{\frac{1}{R_e + \epsilon}} \right) \right] \qquad \forall\, e \in E; L_1, L_2 \in L_d} \qquad (5.13)$$

The following equations determine if the enzyme is being produced $\omega_{eL_1L_2} = 1$ if produced and zero otherwise.

$$\boldsymbol{\omega_{eL_1L_2} \leq VC_{eL_1L_2}} \qquad \forall\ \boldsymbol{e \in E; L_1, L_2 \in L_d} \qquad (5.14)$$

$$\boldsymbol{\omega_{eL_1L_2} \geq \epsilon C_{eL_1L_2}} \qquad \forall\ \boldsymbol{e \in E; L_1, L_2 \in L_d} \qquad (5.15)$$

The following equations, (5.16) and (5.17), determine if the concentration of the enzyme is at sufficient levels (θ_e) to say that the enzyme could be active, $C^+_{eL_1L_2} = 1$ if sufficient concentration, zero otherwise.

$$\boldsymbol{(\theta_e + \epsilon)C^+_{eL_1L_2} \leq C_{eL_1L_2}} \qquad \forall\ \boldsymbol{e \in E; L_1, L_2 \in L_d} \qquad (5.16)$$

$$\boldsymbol{C_{eL_1L_2} \leq \big(V - (\theta_e - \epsilon)\big)C^+_{eL_1L_2} + (\theta_e - \epsilon)} \qquad \forall\ \boldsymbol{e \in E; L_1, L_2 \in L_d} \qquad (5.17)$$

5.6.2.2.5. Enzyme regulation and activity under conditions (equations 5.18 through 5.28)

Determine the net effect of ligands on the enzyme ($\gamma_{eL_dL_{d1}}$) to determine if the protein is active or inactive due to the present ligands ($\delta_{eL_dL_{d1}}$, concentration incorporated through interaction strength $Q_{eL_{d1}}$).

1) Sum of the effects of present ligands and enzymes on the possibility of enzyme e being able to be activated ($\gamma_{eL_dL_{d1}}$), equation (5.18).
2) Determine net effect of activation/inhibition on the enzyme ($\delta_{eL_dL_{d1}}$) equations (5.19) and (5.20).

$$\boldsymbol{\gamma_{eL_1L_2} = \zeta_{\mathrm{e}} + \sum_{e_1 \in E} \left(W_{e_1L_1L_2} B_{ee_1} Q_{ee_1} \right) + B_{eL_1} Q_{eL_1} + B_{eL_2} Q_{eL_2} - B_{eL_1} Q_{eL_1} \sigma_{L_1L_2}} \qquad \forall\, \boldsymbol{e \in E; L_1, L_2 \in L_d} \tag{5.18}$$

$$\boldsymbol{\gamma_{eL_1L_2} \geq -V\left(1 - \gamma^{+}_{eL_1L_2}\right) + \epsilon \gamma^{+}_{eL_1L_2}} \qquad \forall\, \boldsymbol{e \in E; L_1, L_2 \in L_d} \tag{5.19}$$

$$\boldsymbol{\gamma_{eL_1L_2} \leq V\gamma^{+}_{eL_1L_2}} \qquad \forall\, \boldsymbol{e \in E; L_1, L_2 \in L_d} \tag{5.20}$$

Determine if the protein is both produced and can be active. These three constraints, equations (5.21), (5.22), and (5.23), are equivalent to $\kappa_{eL_dL_{d1}} = \omega_{eL_dL_{d1}} \delta_{eL_dL_{d1}}$ (this works because all the variables are binary).

$$\boldsymbol{\kappa_{eL_1L_2} \leq \omega_{eL_1L_2}} \qquad \forall\, \boldsymbol{e \in E; L_1, L_2 \in L_d} \tag{5.21}$$

$$\boldsymbol{\kappa_{eL_1L_2} \leq \gamma^{+}_{eL_1L_2}} \qquad \forall\, \boldsymbol{e \in E; L_1, L_2 \in L_d} \tag{5.22}$$

$$\boldsymbol{\kappa_{eL_1L_2} \geq \omega_{eL_1L_2} + \gamma^{+}_{eL_1L_2} - 1} \qquad \forall\, \boldsymbol{e \in E; L_1, L_2 \in L_d} \tag{5.23}$$

Determine if the protein is produced, active, and at sufficient concentration for it to function. These three constraints, equations (5.24), (5.25), and (5.26), are equivalent to $W_{eL_dL_{d1}} = \kappa_{eL_dL_{d1}} Y_{eL_dL_{d1}}$ (this works because all the variables are binary).

$$\boldsymbol{W_{eL_1L_2} \leq \kappa_{eL_1L_2}} \qquad \forall\, \boldsymbol{e \in E; L_1, L_2 \in L_d} \tag{5.24}$$

$$\boldsymbol{W_{eL_1L_2} \leq C^{+}_{eL_1L_2}} \qquad \forall\, \boldsymbol{e \in E; L_1, L_2 \in L_d} \tag{5.25}$$

$$\boldsymbol{W_{eL_1L_2} \geq \kappa_{eL_1L_2} + C^{+}_{eL_1L_2} - 1} \qquad \forall\, \boldsymbol{e \in E; L_1, L_2 \in L_d} \tag{5.26}$$

Force the logic table to be true in equation (5.27).

$$\boldsymbol{W}_{\boldsymbol{e_d L_1 L_2}} = \boldsymbol{\lambda}_{\boldsymbol{e_d L_1 L_2}} \qquad \forall\, \boldsymbol{e_d} \in \boldsymbol{E_d}; \boldsymbol{L_1}, \boldsymbol{L_2} \in \boldsymbol{L_d} \qquad (5.27)$$

5.6.2.2.6. Attribution of enzyme activity to given conditions under conditions (equations 5.28 through 5.50)

Given all these equations, it is not guaranteed that the circuit produced thus far will truly respond to the input ligands. One persistent issue with the formulation to this point is that a Bistable Orthogonal Design (BOD) can be returned which is independent of the input ligands and the optimization solver will simply choose the appropriate state to appear to meet the logic table. This causes a circuit which appears to the solver to meet design criteria, but in fact does not because it does not respond to ligand conditions. This issue is addressed through what we are choosing to call the attribution constraints. These constraints are created to determine what changes the activity of a protein in a given genetic circuit (e.g. what is the change attributable to?). This is done with several stages of equations.

5.6.2.2.6.1. Set 1: Determine if a particular enzyme pair is encoded (Equations 5.28 through 5.32)

These equations are used to determine if a particular enzyme is encoded (encoded in the binary L_e). This is important in that an enzyme has no attribution from other enzymes and is not attributable to other enzymes.

$$\boldsymbol{L_e \geq \epsilon \sum_{p\in P} \sum_{j\in J} \sum_{t\in T} [M_{pjt}\rho_{je}]} \qquad \forall\, \boldsymbol{e \in E} \qquad (5.28)$$

$$\boldsymbol{L_e \leq V \sum_{p\in P} \sum_{j\in J} \sum_{t\in T} [M_{pjt}\rho_{je}]} \qquad \forall\, \boldsymbol{e \in E} \qquad (5.29)$$

Note that this is formulated as such to allow for multiple transcript copies in a given circuit design. Next, a determination is made as to whether enzyme pairs are encoded (encoded in the binary β_{ee_1}), attribution cannot exist between enzymes.

$$\boldsymbol{\beta_{ee_1} \leq L_e} \qquad \forall\, \boldsymbol{e \in E} \qquad (5.30)$$

$$\boldsymbol{\beta_{ee_1} \leq L_{e_1}} \qquad \forall\, \boldsymbol{e_1 \in E} \qquad (5.31)$$

$$\boldsymbol{\beta_{ee_1} \geq L_e + L_{e_1} + 1} \qquad \forall\, \boldsymbol{e, e_1 \in E} \qquad (5.32)$$

5.6.2.2.6.2. Set 2: Determine if a particular enzyme affects another enzyme's expression (Equations 5.33 through 5.47)

Next, we determine the effect of one enzyme upon the expression of another, through various means. First, through directly affecting enzyme activity (effect of e_1 upon e). Note that the variable D_{ee_1} is restricted to be strictly non-negative.

$$\boldsymbol{D_{ee_1} = |B_{ee_1}|\beta_{ee_1}} \qquad \forall\, \boldsymbol{e, e_1 \in E} \qquad (5.33)$$

Note that the above is linear because B_{ee_1} is a parameter. It was discovered during debugging procedures that attempting to track the sign of attributions can lead to numerical issues (such as an attribution canceling itself out, but still existing); therefore, only the fact of attribution is determined using absolute values. The next group of equations determines the effect of e_1 upon e through controlling the triad expressing e. Note that the variable K_{ee_1} is restricted to be strictly non-negative.

$$\boldsymbol{K_{ee_1} \leq V\beta_{ee_1}} \qquad \forall\, \boldsymbol{e, e_1 \in E} \qquad (5.34)$$

$$\boldsymbol{K_{ee_1} \leq \sum_{p\in P}\sum_{j\in J}\left[|I_{pe_1}| * \sum_{t\in T} M_{pjt}\rho_{je}\right] + V(1 - \beta_{ee_1})} \qquad \forall\, \boldsymbol{e, e_1 \in E} \qquad (5.35)$$

$$\boldsymbol{K_{ee_1} \geq \sum_{p\in P}\sum_{j\in J}\left[|I_{pe_1}| * \sum_{t\in T} M_{pjt}\rho_{je}\right] - V(1 - \beta_{ee_1})} \qquad \forall\, \boldsymbol{e, e_1 \in E} \qquad (5.36)$$

In combination with the domain of K_{ee_1}, $K_{ee_1} = 0$ if $\beta_{ee_1} = 0$, and $K_{ee_1} = \sum_{p\in P}\sum_{j\in J}\left[|I_{pe_1}| * \sum_{t\in T} M_{pjt}\rho_{je}\right]$ otherwise. Next the effect of one enzyme (e_2) upon another enzyme (e) through another enzyme (e_1). This passing of attribution might be through direct enzyme effects (D_{ee_1}) or through the effect of one enzyme upon the triad of another (K_{ee_1}). The variable $\nu'_{e_1e_2}$ below is a binary variable noting if there is attribution of enzyme e_2 upon enzyme e_1 (e.g. e_2 in some way affects the activity of e_1).

$$\boldsymbol{U_{ee_1e_2} \leq V\beta_{ee_1}} \qquad \forall\, \boldsymbol{e, e_1, e_2} \in \boldsymbol{E} \qquad (5.37)$$

$$\boldsymbol{U_{ee_1e_2} \leq \left|B_{ee_1}\right|\nu'_{e_1e_2} + V\left(1 - \beta_{ee_1}\right)} \qquad \forall\, \boldsymbol{e, e_1, e_2} \in \boldsymbol{E} \qquad (5.38)$$

$$\boldsymbol{U_{ee_1e_2} \geq \left|B_{ee_1}\right|\nu'_{e_1e_2} - V\left(1 - \beta_{ee_1}\right)} \qquad \forall\, \boldsymbol{e, e_1, e_2} \in \boldsymbol{E} \qquad (5.39)$$

This can then be condensed into the variable U'_{ee_1} which removes the middle enzyme:

$$\boldsymbol{U'_{ee_1} = \sum_{e_2 \in E} \left[U_{ee_2e_1}\left(1 - \sigma_{ee_1}\sigma_{e_1e_2}\right)\right]} \qquad \forall\, \boldsymbol{e, e_1, e_2} \in \boldsymbol{E} \qquad (5.40)$$

Therefore, U'_{ee_1} represents the indirect attribution of e_1 to the activity of e through direct attributions. This allows any number of intermediates between two enzymes to still count toward attribution due to the effects of networking. Note that the $\left(1 - \sigma_{ee_1}\sigma_{e_1e_2}\right)$ term prevents an enzyme attributing to itself through itself. This prevents a potential self-referential problem which occurs with the definition of ν'_{ee_1}. It should be noted that U'_{ee_1} tracks only enzyme-enzyme interaction networks. Similarly, X_{ee_1} will track enzyme attribution networks through effects on enzyme triads, though due to the need to track triads the formulation is necessarily more complex. Together, U'_{ee_1} and X_{ee_1} allow for full networked tracking of attribution through any number of intermediary enzymes and regulatory mechanisms.

$$\boldsymbol{\chi_{ee_1pjt} \leq VM_{pjt}} \qquad \forall\, \boldsymbol{e, e_1 \in E; p \in P; j \in J; t \in T} \tag{5.41}$$

$$\boldsymbol{\chi_{ee_1pjt} \leq \sum_{e_2 \in E} \left[\left| I_{pe_2} \right| \nu'_{e_2e_1} \rho_{je} \right] + V\left(1 - M_{pjt}\right)} \qquad \forall\, \boldsymbol{e, e_1 \in E; p \in P; j \in J; t \in T} \tag{5.42}$$

$$\boldsymbol{\chi_{ee_1pjt} \geq \sum_{e_2 \in E} \left[\left| I_{pe_2} \right| \nu'_{e_2e_1} \rho_{je} \right] - V\left(1 - M_{pjt}\right)} \qquad \forall\, \boldsymbol{e, e_1 \in E; p \in P; j \in J; t \in T} \tag{5.43}$$

$$\boldsymbol{X_{ee_1} = \sum_{p \in P} \sum_{j \in J} \sum_{t \in T} \left[\chi_{ee_1pjt} \right]} \qquad \forall\, \boldsymbol{e, e_1 \in E} \tag{5.44}$$

Now that the direct (D_{ee_1} and K_{ee_1}) and networked (U'_{ee_1} and X_{ee_1}) attribution variables have been determined, the total attribution can be determined.

$$\boldsymbol{\nu_{ee_1} = D_{ee_1} + K_{ee_1} + U'_{ee_1} + X_{ee_1}} \qquad \forall\, \boldsymbol{e, e_1 \in E} \tag{5.45}$$

Note that ν_{ee_1} is a nonnegative variable, since D_{ee_1}, K_{ee_1}, U'_{ee_1}, and X_{ee_1} are all nonnegative values which may have values greater than 1 depending on the definitions of I_{pa} (for $p \in P$ and $a \in A$) and B_{ea} (for $e \in E$ and $a \in A$). For instance, in some cases it is useful to have values greater than 1 in I_{pa} or B_{ea} to indicate that some effectors are stronger than others. Due to the need for referencing total attribution within the network attribution variables (U'_{ee_1} and X_{ee_1}, which themselves are part of the total attribution) there arises an issue related to the use of multiplication. If a value other than zero or one is used in calculating the total attribution's effect on the network attribution variables, attributions which influence each other could quickly increase in magnitude through recursion. Another potential issue is the possibility that if total attributions are not equal in magnitude, this could result in solution infeasibility as the two attributions cannot exist together. Therefore, there is a need to transform the non-negative ν_{ee_1} into the binary ν'_{ee_1} so that multiplicative identity equations (5.38), (5.39), (5.42), and (5.43) might apply and bypass both these issues. Therefore, ν'_{ee_1} is a binary which is determined using the following constraints.

$$\boldsymbol{\nu_{ee_1} \geq \nu'_{ee_1}} \qquad \forall\, \boldsymbol{e, e_1} \in \boldsymbol{E} \qquad (5.46)$$

$$\boldsymbol{\nu_{ee_1} \leq V\nu'_{ee_1}} \qquad \forall\, \boldsymbol{e, e_1} \in \boldsymbol{E} \qquad (5.47)$$

5.6.2.2.6.3. Set 3: Preventing self-controlling enzymes (equations 5.48 through 5.49)

Now that attribution of one enzyme to another can be determined (ν'_{ee_1}), we have used this variable to prevent an enzyme from directly or indirectly controlling its own expression (which can lead to BODs). This can be prevented by ensuring that there is no self-attribution.

$$\boldsymbol{\nu'_{ee_1} \geq \sigma_{ee_1} - 1} \qquad \forall\, \boldsymbol{e, e_1} \in \boldsymbol{E} \qquad (5.48)$$

$$\boldsymbol{\nu'_{ee_1} \leq 1 - \sigma_{ee_1}} \qquad \forall\, \boldsymbol{e, e_1} \in \boldsymbol{E} \qquad (5.49)$$

5.6.2.2.6.4. Set 4: Prevent the addition of meaningless bioparts (equation 5.50)

The above equations prevent self-attribution and BODs, but do not prevent the addition of meaningless triads to a solution. It was found during development that the addition of meaningless triads was one way for a solution to be reported again at larger circuit sizes. This can be relatively easily fixed with a single equation, which ensures that any encoded enzyme affects circuit reporter enzymes.

$$\boldsymbol{L_e} \leq \sum_{\boldsymbol{e_d \in E_d}} \left[\boldsymbol{v'_{e_d e}}\right] + \boldsymbol{E^{val}_{d,e}} \qquad \forall\, \boldsymbol{e} \in \boldsymbol{E} \qquad (5.50)$$

Where $E^{val}_{d,e} = 1$ if e is a member of the set E_d and $E^{val}_{d,e} = 0$ otherwise. This ensures that each encoded enzyme in some way influences the activity of at least one reporter enzyme or is itself a reporter enzyme.

5.6.2.2.7. Speed Boosting Constraints (equations 5.51 through 5.56)

The following constraints should be implicitly true given all of the previous constraints, yet it was discovered, as with the OptFill tool (Schroeder & Saha, 2020b), that explicitly defining implicit relationships can result in quicker solution times. The following relationship where explicitly defined:

1) Equations (5.51) ensures that all response enzymes are encoded in the genetic circuit.
2) Equations (5.52), (5.53), and (5.54) ensures that no enzyme has activity unless encoded in the genetic circuit.

3) Equations (5.55) and (5.56) ensure that no enzyme has concentration unless encoded in the genetic circuit.

$$\boldsymbol{E_{d,e}^{val} \leq L_e} \qquad \forall\, \boldsymbol{e \in E; L_1, L_2 \in L_d} \qquad (5.51)$$

$$\boldsymbol{W_{eL_1L_2} \leq L_e} \qquad \forall\, \boldsymbol{e \in E; L_1, L_2 \in L_d} \qquad (5.52)$$

$$\boldsymbol{\kappa_{eL_1L_2} \leq L_e} \qquad \forall\, \boldsymbol{e \in E; L_1, L_2 \in L_d} \qquad (5.53)$$

$$\boldsymbol{\omega_{eL_1L_2} \leq L_e} \qquad \forall\, \boldsymbol{e \in E; L_1, L_2 \in L_d} \qquad (5.54)$$

$$\boldsymbol{C_{eL_1L_2} \leq L_e} \qquad \forall\, \boldsymbol{e \in E; L_1, L_2 \in L_d} \qquad (5.55)$$

$$\boldsymbol{C_{eL_1L_2}^{+} \leq L_e} \qquad \forall\, \boldsymbol{e \in E; L_1, L_2 \in L_d} \qquad (5.56)$$

5.6.3. EuGeneCiM Problem Statement and Explanation

While the EuGeneCiM formulation is based upon that of EuGeneCiD, it is markedly less complex due to three factors: i) the design is already known, so M_{pjt} becomes a parameters as opposed to a variable; ii) the design is already complete, attribution need not be tracked; and iii) the transcript an enzyme levels at the current time point are those produced at previous time point(s) and EuGeneCiM is simply solving for the production rate of enzymes and transcripts for the current time point.

5.6.3.1. Objective Function

5.6.3.1.1. Objective Function (Equation 5.57)

The selected objective function is to maximize the production of proteins

$$\boldsymbol{maximize\ Z_M = \sum_{e \in E} \sum_{L_1 \in L_d} \sum_{L_2 \in L_d} \left[C_{eL_1L_2}\right]} \tag{5.57}$$

Note that the objective function is largely unimportant however, as the constraint equations which follow are generally equality constraints, some of which lack variables.

5.6.3.2. Constraint Equations

5.6.3.2.1. Determining the level of transcript production (equations 5.6 through 5.8 and 5.58)

The first set of constraint equations determine the level of transcript production. First, the activity of the promotor under each condition set is evaluated in the same manner as in EuGeneCiD:

$$\boldsymbol{\alpha_{pL_1L_2} = Z_p + \sum_{e \in E}\left[W_{eL_1L_2} I_{pe} H_{pe}\right] + I_{pL_1} H_{pL_1} + I_{pL_2} H_{pL_2} - I_{pL_1} \sigma_{L_1L_2} H_{pL_1} \qquad \forall\, p \in P; L_1, L_2 \in L_d} \tag{5.6}$$

$$\boldsymbol{\alpha_{pL_1L_2} \geq -V\left(1 - \alpha^{+}_{pL_1L_2}\right) + \epsilon \alpha^{+}_{pL_1L_2} \qquad \forall\, p \in P; L_1, L_2 \in L_d} \tag{5.7}$$

$$\boldsymbol{\alpha_{pL_1L_2} \leq V\alpha^{+}_{pL_1L_2}} \qquad \forall\, \boldsymbol{p \in P; L_1, L_2 \in L_d} \tag{5.8}$$

Then, the level of transcript production under each condition can be evaluated, similar to as is done in equations (5.9) through (5.12) with two distinct simplifications: i) as M_{pjt} is a parameter, the linearization of $S_p M_{pjt} \alpha^{+}_{pL_1L_2}$ accomplished in equations (5.9) through (5.11) is no longer needed, and is substituted directly into equation (5.12) and ii) degradation of RNA is handled in another programmatic step between the time points, rather than at a single time point as in EuGeneCiD, therefore this is not included.

$$\boldsymbol{\phi^{prod}_{jtL_1L_2} = \sum_{p \in P}\left[\left(S_p M_{pjt} \alpha^{+}_{pL_1L_2} + M_{pjt} F_p\right)\right]} \qquad \forall\, \boldsymbol{j \in J; t \in T; L_1, L_2 \in L_d} \tag{5.58}$$

Note that the superscript $prod$ is added to $\phi^{prod}_{jtL_1L_2}$ to indicated that this is the transcript production at the current time point. This is an important distinction as the transcript carried over from the previous time point is denoted $\phi^{t_{n-1}}_{jtL_1L_2}$ and is used to calculate the protein production at time t_n. This arrangement allows for the simulation of the delay between triad activation and transcript production, as well as between transcript production and enzyme expression. Also, note that the identity of the terminator is tracked in $\phi^{t_n}_{jtL_1L_2}$ as the terminator determines the half-life of its associated transcript.

5.6.3.2.2. Determining the level of protein production (equation 5.59)

As mentioned, the amount of protein produced at time t_n is calculated from the amount of transcript carried over from the previous time point t_{n-1}. This is calculated in the following equation, which is analogous to equation (5.13) without the degradation term.

$$\boldsymbol{C^{prod}_{eL_1L_2}} = \sum_{\boldsymbol{j \in J}} \left[\boldsymbol{\rho_{je} \eta_j} \sum_{\boldsymbol{t \in T}} \boldsymbol{\phi^{t_{n-1}}_{jtL_1L_2}} \right] \qquad \forall\, \boldsymbol{e \in E; L_1, L_2 \in L_d} \qquad (5.59)$$

Note that $C^{prod}_{eL_1L_2}$ represents to protein production at time t_n, and that the activity of those proteins is determined by the carry-over from the previous time point, $C^{t_{n-1}}_{eL_1L_2}$.

5.6.3.2.3. Determining the activity of the proteins (equations 5.18 through 5.26)

Using the carry-over protein concentration, $C^{t_{n-1}}_{eL_1L_2}$, the activity of the enzyme is calculated in the same way as in EuGeneCiD and utilizing the same equations. These equations are restated here, see the symbols used section for symbol definitions.

$$\begin{aligned} \boldsymbol{\gamma_{eL_1L_2}} = \boldsymbol{\zeta_{\mathrm{e}}} + \sum_{\boldsymbol{e_1 \in E}} \left(\boldsymbol{W_{e_1L_1L_2} B_{ee_1} Q_{ee_1}} \right) & \\ + \boldsymbol{B_{eL_1} Q_{eL_1}} + \boldsymbol{B_{eL_2} Q_{eL_2}} & \qquad \forall\, \boldsymbol{e \in E; L_1, L_2 \in L_d} \qquad (5.18) \\ - \boldsymbol{B_{eL_1} Q_{eL_1} \sigma_{L_1L_2}} & \end{aligned}$$

$$\boldsymbol{\gamma_{eL_1L_2}} \geq -\boldsymbol{V}\left(\boldsymbol{1} - \boldsymbol{\gamma^+_{eL_1L_2}}\right) + \boldsymbol{\epsilon \gamma^+_{eL_1L_2}} \qquad \forall\, \boldsymbol{e \in E; L_1, L_2 \in L_d} \qquad (5.19)$$

$$\boldsymbol{\gamma_{eL_1L_2} \leq V\gamma^{+}_{eL_1L_2}} \qquad \forall\, \boldsymbol{e \in E; L_1, L_2 \in L_d} \tag{5.20}$$

$$\boldsymbol{\kappa_{eL_1L_2} \leq \omega_{eL_1L_2}} \qquad \forall\, \boldsymbol{e \in E; L_1, L_2 \in L_d} \tag{5.21}$$

$$\boldsymbol{\kappa_{eL_1L_2} \leq \gamma^{+}_{eL_1L_2}} \qquad \forall\, \boldsymbol{e \in E; L_1, L_2 \in L_d} \tag{5.22}$$

$$\boldsymbol{\kappa_{eL_1L_2} \geq \omega_{eL_1L_2} + \gamma^{+}_{eL_1L_2} - 1} \qquad \forall\, \boldsymbol{e \in E; L_1, L_2 \in L_d} \tag{5.23}$$

$$\boldsymbol{W_{eL_1L_2} \leq \kappa_{eL_1L_2}} \qquad \forall\, \boldsymbol{e \in E; L_1, L_2 \in L_d} \tag{5.24}$$

$$\boldsymbol{W_{eL_1L_2} \leq C^{+}_{eL_1L_2}} \qquad \forall\, \boldsymbol{e \in E; L_1, L_2 \in L_d} \tag{5.25}$$

$$\boldsymbol{W_{eL_1L_2} \geq \kappa_{eL_1L_2} + C^{+}_{eL_1L_2} - 1} \qquad \forall\, \boldsymbol{e \in E; L_1, L_2 \in L_d} \tag{5.26}$$

5.6.3.2.4. Modeling degradation of transcripts and enzyme

Between time points, and attempted solutions of EuGeneCiM, degradation of the bioparts are calculated as follows:

$$\boldsymbol{\phi^{t_n}_{jtL_1L_2} = \left(\phi^{prod}_{jtL_1L_2} + \phi^{t_{n-1}}_{jtL_1L_2}\right)\left(0.5^{\left(\frac{1}{G_t/1+\epsilon}\right)}\right)} \qquad \forall\, \boldsymbol{j \in J; t \in T; L_1, L_2 \in L_d} \tag{5.60}$$

$$\boldsymbol{C^{t_n}_{eL_1L_2} = \left(C^{prod}_{eL_1L_2} + C^{t_{n-1}}_{eL_1L_2}\right)\left(0.5^{\left(\frac{1}{R_e/2+\epsilon}\right)}\right)} \qquad \forall\, \boldsymbol{e \in E; L_1, L_2 \in L_d} \tag{5.61}$$

Note that there is one major difference in the degradation terms of equations (5.60) and (5.61): the half-lives are reduced by half in EuGeneCiM compared to EuGeneCiD. This in attempt to reconcile the differences between EuGeneCiD and EuGeneCiM when considering the cumulative effects of dynamic modeling. This occurs because, while EuGeneCiD accounts for a single time point and EuGeneCiM accounts for several, the enzyme and transcript accumulations in EuGeneCiM were generally one or two order of magnitude higher than that predicted in EuGeneCiD. This was an issue because the same concentration thresholds existed for enzyme activity, and therefore resulted in no enzyme being in an "off" state after sufficient time in EuGeneCiM. This fix reduces the half-life of transcripts and enzymes, resulting in closer parity in concentration and modeling of circuit designs while minimizing the number of parameters perturbed.

5.6.3.2.5. Other important aspects of EuGeneCiM formulation

Constraints not included in the formulation can be as important as those which are and can serve to highlight the function of the problem. Specifically, no constraints are included which force the provided conceptualization (in the form of a logic table) to be true. This is for two reasons. The first is that, in solving in a point by point manner, there will inevitably be time points in which the logic table is not true, particularly due to the delays between transcription and translation built into the tool. Secondly, this allows EuGeneCiM to be a screening process to remove any designs which function differently when no longer optimizing for desired behavior or when considering dynamic behavior.

5.6.4. Designing and Modeling Genetic Circuits

See Figure 5.4 for a visual representation of the overall workflow and to specifically illustrate how the EuGeneCiD and EuGeneCiM formulations fit into this workflow. This work began with the conceptualization of synthetic biology interventions. For the purposes of demonstrating these design and modeling tools, simple circuit conceptualizations were selected, namely the two input circuits of AND, NIMPLY, HALF ADDER, NAND, NOR, XNOR, and XOR. Note that logic gates will be capitalized throughout this text to avoid confusion. These particular conceptualizations were chosen because they are easy to represent in logic table format, and well-known, and often studied in the context of genetic circuits (particularly NOR and NIMPLY) (Borujeni et al., 2020)(Tan & Ng, 2021). A library of bioparts (consisting of promotors, transcripts, terminators, and proteins) was then selected which were i) native to Arabidopsis (particularly promotors), ii) demonstrated to be functional in synthetic biology applications in Arabidopsis, or iii) were from related plant species which we judged were likely to function in synthetic biology applications. Note that when a particular biopart had different expression or regulation patterns at different stages in growth or in different tissues, the pattern related to seedling root was selected. These parts are described in detail in Supplemental Table S2 These two items, conceptualizations and the bioparts library, are then appropriately formatted as input files utilizing a Perl script (included in the associated GitHub at github.com/ssbio/EuGeneCiDM) which reads a dababase filed appended with the desired circuit logic, example is provided in Supplemental Table 3 with the full set used here in the associated GitHub at github.com/ssbio/EuGeneCiDM), and writes the input files accordingly. EuGeneCiD was implemented in the Generalized Algebraic Modeling System (GAMS) language and run using the CPLEX solver. At this point, the workflow diverts to several possible outcomes. First, EuGeneCiD found no designs of the appropriate size, indicated by no solution or an “integer infeasible” model status. This is addressed by incrementing the allowed model size by one, provided the maximum allowable circuit design size has not been exceeded, and re-attempting to solve EuGeneCiD. The second possibility is that EuGeneCiD found a potential design which fits the current criteria. This design will be the output of EuGeneCiD and the input of EuGeneCiM. EuGeneCiM then simulates the designed circuit, beginning at time point zero with no initial concentration of any enzyme or transcript. EuGeneCiM will return, as an output, the relative production of enzymes and transcripts at the given time point. The concentration of enzymes at the current time point is reduced according to the half-life characteristics of the enzyme or transcript terminator, and the newly produced amount of each is added to this value as the carry over to the next time point. EuGeneCiM is then solved for the next time point, and the process is repeated until all time points have a solution. From this, the dynamic behavior of the designed circuit may be plotted as a visual representation of the circuit simulation. This can be done through an additional Perl script (included in the associated GitHub at github.com/ssbio/EuGeneCiDM). The cycle of design (through EuGeneCiD) and simulation (through EuGeneCiM) continues until case two occurs. The final possible outcome of EuGeneCiD is that no designs of the appropriate size can be found, and that incrementing the size would result in exceeding the maximum allowable circuit design size (here, ten triads). In this case, it will be concluded that there are no further designs, and the design and simulation results should be manually screened to pick the most promising design candidate(s). The example given here is a set of Cd/Cu responsive AND circuit from which is selected solution #41, which has the highest objective value.

5.6.5. Computing, Language, and Solving Resources in Implementation

This study has produced several unique software codes in the form of GAMS or Perl programming languages/tools. For implementing and solving EuGeneCiD and EuGeneCiM the Generalized Algebraic Modeling System (GAMS), version 24.7.4 was used in conjunction with the CPLEX solver version 12.6. Scripts which automate certain tasks utilize Perl version 5.26 for Unix or Strawberry Perl 5.24.0.1 for Windows. The code provided is compatible with both versions. The main workflow (previously described) was implemented on the Holland Computing Center Crane Cluster and allowed to run for at most seven days (168 hours) before being terminated. CPLEX solver settings used are included in the associated GitHub at github.com/ssbio/EuGeneCiDM.

5.6.6. Quantification and Statistical Analysis

Many values used in the definition of the bioparts in the database used were defined through manual quantification of quantitative data. For promotors, normal state was determined by literature evidence (either normally on or off). Strength and leakiness were determined, when possible, from western or northern blot images, with strong expression being given a value of five and no expression being given a value of one. In some cases, the fold change in expression of a gene associated with a given promotor was known under induced cases. In these cases, the ratio or strength to leakiness was adjusted to reflect these known expression changes. Inducer and repressor identities were identified using literature evidence. The base strength of induction or repression was set to one; however, if some ligand showed greater activation or repression than another, a value of two was assigned to model a greater effect on the activity of that particular promotor. For transcripts, the transcriptional efficiency can represent various design elements of the gene, codon optimization for instance, which can change the speed or efficiency of translation of the gene. A value of zero would indicate that the gene cannot be translated and a value of three would indicate efficient translation. In this work, there was no such adjusting of the translational properties of the genes; therefore, a base value of two was assumed for all translational efficiencies. A small set of three terminators were identified from Nagaya *et al.*, 2010 and the relative half-lives of these terminators were determined as follows. The scale used was from zero representing near instant of mRNA to three representing slow degradation of associated mRNA. Based on Nagaya *et al.*, 2010 values of associated mRNA half-life for each terminator was quantified. For enzymes, the default state was determined from literature. The default expression and half-life were assumed to be five and two, respectively. These values were changed if literature evidence was found to warrant the need to adjust these values. For instance, cI was noted as being rapidly degraded in registry of standard biological parts, and therefore given a shorter half-life.

Chapter 6

6. CONCLUSIONS AND GOING FORWARD

Portions of this material have previously appeared in the following publication:

W. L. Schroeder, R. Saha, OptFill: A Tool for Infeasible Cycle-Free Gapfilling of Stoichiometric Metabolic Models, iScience, 23(2020) 1-14. Used with permission.

W. L. Schroeder, S. D. Harris, and R. Saha, Computation-Driven Analysis of Model Polyextremotolerant Fungus Exophiala dermatitidis: Defensive Pigment Metabolic Costs and Human Applications, iScience, 23(2020) 1-17. Used with permission.

W. L. Schroeder, R. Saha, Introducing an Optimization- and explicit Runge-Kutta- based Approach to Perform Dynamic Flux Balance Analysis, Scientific Reports, 10:9241(2020) 1-28. Used with permission.

W. L. Schroeder, R. Saha, Protocol for Genome-Scale Reconstruction and Melanogenesis Analysis of Exophiala dermatitidis, STAR Protocols, 1(2020) 1-37. Used with permission.

W. L. Schroeder, R. Saha, Protocol for Genome-Scale Reconstruction and Melanogenesis Analysis of Exophiala dermatitidis, STAR Protocols, 1(2020) 1-37. Used with permission.

M. M. Islam, W. L. Schroeder, and R. Saha, Kinetic Modeling of Metabolism: Present and Future, Current Opinion in System Biology, (2021) 1-7. Used with permission.

6.1. PREFACE

In this dissertation, I have detailed several unique optimization-based tools, including OptFill, the TIC-Finding Problem, ORKA, EuGeneCiD, and EuGeneCiM, which apply to systems biology and its application via synthetic biology. In this section, I briefly summarize each work and its implications for the field of systems biology. This chapter is concluded by discussing the future of the most promising research project, as well as research areas I would like to apply my expertise to in future.

6.2. SUMMARY AND CONCLUSIONS

In Chapter 2, the OptFill tool is introduced. In this chapter, the tool is developed, its mathematical formulation is detailed, and it is applied to three test models and one GSM of *E. coli*. It is noted that the initial published formulation has issues with speed and computational tractability. In Chapter 3, OptFill is revisited, with a revised and more tractable formulation which is applied to the GSM reconstruction of the highly melanized fungi *E. dermatitidis*. With this revised formulation, OptFill can serve an important role in future genome-scale modeling efforts. First, as a holistic (that is, solving on a whole-model basis) and conservative (that is, minimizing the number of reactions added to the reconstruction) reconstruction tool, OptFill serves a different function than other gapfilling tools which work on a per-metabolite basis and which is subject to the curator's approach. Secondly, the modified TIC-Finding Problem (mTFP) discussed in Chapter 2 provides a unique and valuable model curation tool for the identification of TICs for which no other robust tool is available. It is hoped that various automated GSM reconstruction tools, such as ModelSeed, KBase, or CarveMe, or individuals manually reconstructing or curating GSM will soon incorporate OptFill or the mTFP into their model reconstruction workflows. To this end a protocol detailing how to use OptFill and the mTFP has been published in STAR Protocols to encourage their use.

Chapter 3 introduces the first GSM reconstruction of the metabolism of *E. dermatitidis*, accomplished using the OptFill tool. Aside from the previously discussed improvements upon OptFill, this chapter detailed the cost, in terms of shadow price, to *E. dermatitidis* for the production of melanin and carotenoid defensive pigments. This analysis led to two interesting hypothesis from the observations that the shadow price of carotenoids is greater than that of melanins and the observation that shadow price varies based on nutrients in the media. The first hypothesis is that carotenoids play an important though undiscovered role in *E. dermatitidis*. This comes from the fact that it was noted that there is no function which (the more expensive) carotenoids can perform which melanins cannot also perform and that melanins are the first line of defense, being deposited in the cell wall. The second hypothesis is that *E. dermatitidis* produces such a wide array of defensive pigments so that it may attempt to minimize the cost of its defense through its defensive pigment array. The second investigation in Chapter 3 was to study if *E. dermatitidis* could serve as a model melanin-producing organism, particularly for human melanocytes. It was concluded that, due to the strong conservation of tyrosinase active site residues and similarity in eumelanin synthesis pathways, *E. dermatitidis* could serve as a model of human melanocytes. This is important an exciting for the future research directions which this research could take (see Section 6.3).

Chapter 4 introduces a new approach to modeling time-dependent metabolism through dFBA, namely the Optimization- and Runge-Kutta- based Approach (ORKA). This chapter then demonstrates the application of ORKA to the model plant system *Arabidopsis thaliana*, in a model which spans its lifetime (from 0 to 61 days after germination), models four distinct tissues (leaf, seed, stem, and root), and captures *in vivo* plant-level behavior in an *in silico* model. This is a considerable improvement on the previous best dFBA model of *A. thaliana*, which modeled two tissues from 6 to 36 days after germination, did not capture *in vivo* plant-scale behavior, and had a considerably higher error magnitude in solving ordinary differential equations (about 5000x greater). The ORKA then can be applied to accurately model metabolism across time. Further, as the formulation of the ORKA is largely generalized, particularly for the Runge-Kutta method used, modelers using this approach in future can choose a method which provides their desired balance of tractability, speed, and numerical stability.

In Chapter 5, the Eukaryotic Genetic Circuit Design (EuGeneCiD) and Modeling (EuGeneCiM) tools were introduced. These tools use optimization and mixed-integer linear programming to design and model synthetic biology applications, to increase the chances of implementing a successful design. In this chapter, EuGeneCiD and EuGeneCiM are applied to the task of designing and modeling 27 unique genetic circuits which respond to divalent metal ions in *Arabidopsis thaliana*. Additionally, it is demonstrated that EuGeneCiM may be used as a stand-alone tool which can model complex synthetic biology applications, such as repressilators. These tools, utilizing optimization, allow for the identification of optimal and non-intuitive circuit designs and can help move the field of synthetic biology away from intuitive designs to computational designs. Ideally, these tools will be incorporated into the synthetic biology application workflows of researchers or research groups.

6.3. GOING FORWARD

The research detailed here focuses on Genome-Scale Modeling and associated techniques. These models have proven informative and accurate in many applications, and recent advances in computational techniques, computing, genome annotation, and high-throughput biology have led to an ever-increasing number of GSMs for a wider and more diverse array of organisms. As much of this research is the development of mathematic-based tools, it is hoped that these tools may be incorporated into automated GSM reconstruction workflows in some capacity such as exist in ModelSeed and KBase.

Perhaps the research project described here with the greatest future potential is the modeling of *Exophiala dermatitidis* (Chapter 3). At this time, the two primary hypotheses generated by this work, that carotenoids have an as yet undiscovered function and that *E. dermatitidis* engages in phenotype cost minimization, are now being investigated *in vivo* by our collaborator Dr. Steven D. Harris. Further, there a several additional possible studies which spring from this work. The first potential follow-up study is to further investigate the metabolic similarity between human melanocytes and *E. dermatitidis* with respect to eumelanin and pheomelanin synthesis. This study could be an *in silico* investigation to further determine the suitability of *E. dermatitidis* as a model of a human melanocyte. This comparison can be done using the latest human GSM (Robinson et al., 2020) and specializing this model to create a melanocyte-specific GSM. This study would encompass studying metabolic flux ranges and variability; shadow cost; gene expression under various stimuli conditions; and metabolic reprogramming associated with different types of albinism (particularly oculocutaneous albinism type 1, OCA1). Should this study suggest that *E. dermatitidis* would make a suitable model system, collaborations with *in vivo* researchers might be established to identify potential albinism treatment options, which could have enormous medical and social impact in Africa (Brilliant, 2015). Otherwise, should *E. dermatitidis* be shown as a poor model system for human melanocytes, this leads to the possibility of *E. dermatitidis* as a model polyextremotolerant organism, which may be useful for redesigning organism for harsh conditions, such as for post-climate change or extraterrestrial environments.

In future, I would also like to extend my research to other types of models. GSMs are exclusively studied here; however, these models are limited by their reliance on linear relationships between modeled fluxes, frequent reliance on the Pseudo-Steady State Hypothesis (PSSH), their lack of enzyme kinetics, and difficulty of incorporating enzyme regulatory mechanisms. Particularly promising is the emergence of Kinetic Metabolic Models (KiMMs), which incorporate metabolite concentration, enzyme regulation, and enzyme kinetics into metabolic modeling. At present, these models have been limited due to computational expense, lack of biological knowledge, and difficulty in parameterization. This is reflected in the relative size of kinetic models compared to other modeling approaches. For example, the largest KiMM of which I am aware of for *E. coli*, the k-ecoli457 model, has only 457 reactions [31]. This is an order of magnitude smaller than other types of available E .coli models such as iJL1678-ME [56], the more recent iML1515 model [57], or the earlier iJO1366 [58]. However, due to advances in computing and high-throughput biological techniques, these types of models are becoming increasingly more feasible, less limited, and more accurate. I believe it likely that, in future, KiMMs will supplant GSMs as the standard systems biology model. This usurpation will be due to several factors, including the increase in computing technology, the ability to model more complex phenomena (such as pharmacokinetics or cheminformatics), and the increased accuracy which results from incorporating enzyme kinetics. Therefore, extending my future research expertise to this model type will be important to my research remaining current and competitive.

Chapter 7

7. APPENDIX

7.1. PREFACE

This chapter will detail various important information relevant to this dissertation which is not detailed elsewhere. This includes information on how to access supporting and supplemental files and a list of acronyms used throughout this text.

7.2 SUPPLEMENTAL AND SUPPORTING FILES

Given the nature of this research, especially the required file architecture for the GAMS programming language, it is impractical to include in this dissertation all code which is created to accomplish, utilize, or facilitate this research. Therefore, all necessary code to replicate this research or to use these tools in another context is available in our research group GitHub at github.com/ssbio/. Repositories have been made corresponding to specific chapter. For Chapter 2, the appropriate repository is located at github.com/ssbio/OptFill. For Chapter 3, the appropriate repository is located at github.com/ssbio/E_dermatitidis_model. For Chapter 4, the appropriate repository is located at github.com/ssbio/p-ath773. For Chapter 5, the appropriate repository is located at github.com/ssbio/EuGeneCiD. Further, since many of the supplemental files are large Microsoft Excel-based tables or large Microsoft Word files, our supplemental and supporting files are also made available through GitHub. These can be found at github.com/ssbio/OptFill/Supplementals (Chapter 2), github.com/ssbio/E_dermatitidis_model/Supplemental_Files (Chapter 3), github.com/ssbio/path773/Supplemental_Files (Chapter 4), and github.com/ssbio/EuGeneCiDM/Supplemental_Files (Chapter 5).

7.3 EFFORTS FOR STUDY RELICABILITY

While not discussed in the previous chapters, efforts have been made to increase the replicability of this work and encourage other researchers to make use of these tools. This includes making all supplemental and supporting files available through public GitHub repositories as well as submitting and publishing protocol articles to guide others in the use of these tools. At present, we have one protocol published in (related to Chapters 2 and 3) and one protocol submitted to (related to Chapter 5) the journal STAR Protocols. This is an effort both to enhance the replicability of our studies, as well as encourage other researchers to use the tools and models which I have developed.

7.4. ACRONYMS USED

Below is a list of acronyms which are used throughout the text.

LHS – Left Hand Side

RHS – Right Hand Side

TFP – TIC-Finding Problem

CPs – Connecting Problems

FBA – Flux Balance Analysis

TM1 – First Test Model in OptFill study

TDb1 – First Test Database in OptFill study

TM2 – Second Test Model in OptFill study

TDb2 – First Test Database in OptFill study

TM3 – Third Test Model in OptFill study

TDb3 – First Test Database in OptFill study

GAM – Growth Associated Maintenance

NGAM – Non-Growth Associated Maintenance

GPR – Gene-Protein-Reaction

SM – Stoichiometric Model

GSM – Genome-Scale Model

GEM – GEnome-scale Model

FBA – Flux Balance Analysis

dFBA – dynamic Flux Balance Analysis

FVA – Flux Variability Analysis

LP – Linear Problem

MILP – Mixed Integer Linear Problem

Arabidopsis – Arabidopsis thaliana

wrt – with respect to

gDW – grams Dry Weight

DW – Dry Weight

gFW – grams Fresh Weight

FW – Fresh Weight

MFA – Metabolic Flux Analysis

KEGG – Kyoto Encyclopedia of Genes and Genomes

DAG – Days After Germination

HAG – Hours After Germination

COBRA – COnstraint-Based Reconstruction and Analysis

SOA – Static Optimization-based dFBA Approach

DOA – Dynamic Optimization-based dFBA Approach

ORKA – Optimization- and explicit Runge-Kutta dFBA Approach

HAG – Hours After Germination

PSSH – Pseudo-Steady State Hypothesis

GAMS – Generalized Algebraic Modeling System

TFP – TIC-Finding Problem (part of OptFill)

TIC – Thermodynamically Infeasible Cycle

CP – Connecting Problem (part of OptFill)

7.5 CANDIDATE PUBLICATION LIST

In partial fulfilment of the Doctorate of Philosophy degree requirements for the Chemical Engineering department, I have published five peer-reviewed journal articles in prestigious scientific journals. For all of these publications, I am either the first author or co-first author (first author denoted by *). These articles are listed below.

Wheaton L. Schroeder* and Rajib Saha. "Protocol for Genome-Scale Reconstruction and Melanogenesis Analysis of *Exophiala dermatitidis*". *STAR Protocol*, vol. 1, Sept. 18, 2020. Available: https://star-protocols.cell.com/protocols/183 (doi: https://doi.org/10.1016/j.xpro.2020.100105). Journal impact factor: not determined.

Wheaton L. Schroeder* and Rajib Saha. "Introducing an optimization- and explicit Runge-Kutta-based approach to perform dynamic flux balance analysis". *Scientific Reports*, vol. 10, no. 9241, Jun. 8, 2020. Available: https://www.nature.com/articles/s41598-020-65457-4 (doi: https://doi.org/10.1038/s41598-020-65457-4). Journal impact factor: 4.2.

Wheaton L. Schroeder*, Steven D. Harris, and Rajib Saha. "Computation-driven analysis of model polyextremotolerant fungus *Exophiala dermatitidis*: defensive pigment metabolic costs and human applications". *iScience*, vol. 23 no. 4, Apr. 24, 2020. Available: https://www.cell.com/iscience/fulltext/S2589-0042(20)30164-4 (doi: https://doi.org/10.1016/j.isci/2020.100980). Journal impact factor: 4.4.

Wheaton L. Schroeder and Rajib Saha. "OptFill: a tool for infeasible cycle-free gapfilling of stoichiometric metabolic models". *iScience,* vol. 23 no. 1, pp. 1-14, Jan. 24, 2020. Available: https://www.cell.com/iscience/fulltext/S2589-0042(19)30528-0 (doi: https://doi.org/10.1016/j.isci.
2019.100783). Journal impact factor: 4.4.

Mohammad Mazharul Islam*, Wheaton Lane Schroeder*, and Rajib Saha. "Kinetic Modeling of Metabolism: Present and Future" (*Invited Review*, Current Opinion in Systems Biology, accepted for publication, anticipated publication by May 2021). Journal pre-proof available at doi: https://doi.org/10.1016/j.coisb.2021.04.003.

I have further submitted thee additional manuscripts for peer review in prestigious journals. In all these works, I am either first author or co-first author (first author denoted by *). These articles are listed below.

Wheaton L. Schroeder[*], Anna Baber, and Rajib Saha. "Optimization-based Eukaryotic Genetic Circuit Design (EuGeneCiD) and Modeling (EuGeneCiM) Tools: Computational Approach to Synthetic Biology" (Submitted to *iScience*, anticipated publication by Aug. 2021).

Wheaton L. Schroeder[*], Anna Baber, and Rajib Saha. "Protocol for the use of Eukaryotic Genetic Circuit Design (EuGeneCiD) and Modeling (EuGeneCiM) Tools" (Submitted to *STAR Protocols*, anticipated publication by Aug. 2021).

Niaz Bahar Chowdhury[*], Wheaton L. Schroeder[*], Debolina Sarkar, Nardjis Amiour, Isabelle Quilleré, Bertrand Hirel, Costas D. Maranas, and Rajib Saha. "Dissecting the Metabolic Reprogramming of Maize Root under Nitrogen Limiting Stress Condition" (Submitted to *New Phytologist*, anticipated publication by Aug. 2021).

Chapter 8

8. REFERENCES

Alsiyabi, A., Immethun, C. M., & Saha, R. (2019). Modeling the Interplay between Photosynthesis, CO2 Fixation, and the Quinone Pool in a Purple Non-Sulfur Bacterium. *Scientific Reports*, *9*(1), 1–9. https://doi.org/10.1038/s41598-019-49079-z

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402. https://doi.org/10.1093/nar/25.17.3389

Altschul, S. F., Wootton, J. C., Gertz, E. M., Agarwala, R., Morgulis, A., Schäffer, A. A., & Yu, Y. K. (2005). Protein database searches using compositionally adjusted substitution matrices. *FEBS Journal*, *272*(20), 5101–5109. https://doi.org/10.1111/j.1742-4658.2005.04945.x

Andersen, M. R., Nielsen, M. L., & Nielsen, J. (2008). Metabolic model integration of the bibliome, genome, metabolome and reactome of Aspergillus niger. *Molecular Systems Biology*, *4*(178), 178. https://doi.org/10.1038/msb.2008.12

Arkin, A. P., Cottingham, R. W., Henry, C. S., Harris, N. L., Stevens, R. L., Maslov, S., … Yu, D. (2018). KBase: The United States department of energy systems biology knowledgebase. *Nature Biotechnology*, *36*(7), 566–569. https://doi.org/10.1038/nbt.4163

Avalos, J., & Carmen Limón, M. (2015). Biological roles of fungal carotenoids. *Current Genetics*, *61*(3), 309–324. https://doi.org/10.1007/s00294-014-0454-x

Baleja, R., Sumpich, J., Bos, P., Helstynova, B., Sokansky, K., & Novak, T. (2015). Comparison of LED properties, compact fluorescent bulbs and bulbs in residential areas. *Proceedings of the 2015 16th International Scientific Conference on Electric Power Engineering, EPE 2015*, 566–571. https://doi.org/10.1109/EPE.2015.7161181

Baud, S., Boutin, J., Miquel, M., Lepiniec, L., & Rochat, C. (2002). An integrated overview of seed development in Arabidopsis thaliana ecotype WS. *Plant Physiology Biochemistry*, *40*, 151–160.

Beyer, P., Al-Babili, S., Ye, X., Lucca, P., Schaub, P., Welsch, R., & Potrykus, I. (2002). Golden Rice: introducing the beta-carotene biosynthesis pathway into rice endosperm by genetic engineering to defeat vitamin A deficiency. *The Journal of Nutrition*, *132*(3), 506S-510S. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11880581

Bordbar, A., Yurkovich, J. T., Paglia, G., Rolfsson, O., Sigurjónsson, Ó. E., & Palsson, B. O. (2017). Elucidating dynamic metabolic physiology through network integration of quantitative time-course metabolomics. *Scientific Reports*, *7*(March), 1–12. https://doi.org/10.1038/srep46249

Borujeni, A. E., Zhang, J., Doosthosseini, H., Nielsen, A. A. K., & Voigt, C. A. (2020). Genetic circuit characterization by inferring RNA polymerase movement and ribosome usage. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-18630-2

Boyes, D. C., Zayed, A. M., Ascenzi, R., Mccaskill, A. J., Hoffman, N. E., Davis, K. R., & Görlach, J. (2001). Growth Stage-Based Phenotypic Analysis of Arabidopsis : A Model for High Throughput Functional Genomics in Plants. *The Plant Cell*, *13*(July), 1499–1510.

Brilliant, M. H. (2015). Albinism in Africa: A medical and social emergency. *International Health*, *7*(4), 223–225. https://doi.org/10.1093/inthealth/ihv039

Brunk, E., Sahoo, S., Zielinski, D. C., Altunkaya, A., Dräger, A., Mih, N., … Palsson, B. O. (2018). Recon3D enables a three-dimensional view of gene variation in human metabolism. *Nature Biotechnology*, *36*(3), 272–281. https://doi.org/10.1038/nbt.4072

Burgard, A. P., Nikolaev, E. V., Schilling, C. H., & Maranas, C. D. (2004). Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome Research*, *14*(2), 301–312. https://doi.org/10.1101/gr.1926504

Burgard, A. P., Pharkya, P., & Maranas, C. D. (2003). OptKnock: A Bilevel Programming

Framework for Identifying Gene Knockout Strategies for Microbial Strain Optimization. *Biotechnology and Bioengineering*, *84*(6), 647–657. https://doi.org/10.1002/bit.10803

Camacho, E., Vij, R., Chrissian, C., Prados-Rosales, R., Gil, D., O'Meally, R. N., … Casadevall, A. (2019). The structural unit of melanin in the cell wall of the fungal pathogen Cryptococcus neoformans. *Journal of Biological Chemistry*, *294*(27), 10471–10489. https://doi.org/10.1074/jbc.RA119.008684

Cannell, M. G. R., & Thornley, J. H. M. (1999). Modeling the Components of Plant Respiration: Some Guiding Principles. *Annals of Botany*, *85*, 45–54.

Caspi, R. (2006). MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Research*, *34*(90001), D511–D516. https://doi.org/10.1093/nar/gkj128

Caspi, Ron, Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C. A., … Karp, P. D. (2014). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research*, *42*(D1), 459–471. https://doi.org/10.1093/nar/gkt1103

Chan, S. H. J., Wang, L., Dash, S., & Maranas, C. D. (2018). Accelerating flux balance calculations in genome-scale metabolic models by localizing the application of loopless constraints. *Bioinformatics (Oxford, England)*, *34*(24), 4248–4255. https://doi.org/10.1093/bioinformatics/bty446

Chen, T. H. H., & Murata, N. (2002). Enhancement of tolerance of abiotic stress by metabolic engineering of betaines and other compatible solutes. *Current Opinion in Plant Biology*, *5*(3), 250–257. https://doi.org/10.1016/S1369-5266(02)00255-8

Chen, Y., Zhang, S., Young, E. M., Jones, T. S., Densmore, D., & Voigt, C. A. (2020). Genetic circuit design automation for yeast. *Nature Microbiology*, *5*(November). https://doi.org/10.1038/s41564-020-0757-2

Chen, Z., Martinez, D. A., Gujja, S., Sykes, S. M., Zeng, Q., Szaniszlo, P. J., … Cuomo, C. A. (2014). Comparative genomic and transcriptomic analysis of Wangiella dermatitidis, a major cause of phaeohyphomycosis and a model black yeast human pathogen. *G3*, *4*(4), 561–578. https://doi.org/10.1534/g3.113.009241

Cheung, C. Y. M., Williams, T. C. R., Poolman, M. G., Fell, D. A., Ratcliffe, R. G., & Sweetlove, L. J. (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*(6), 1050–1061. https://doi.org/10.1111/tpj.12252

Chowdhury, R., Chowdhury, A., & Maranas, C. D. (2015). Using gene essentiality and synthetic lethality information to correct yeast and CHO cell genome-scale models. *Metabolites*, *5*(4), 536–570. https://doi.org/10.3390/metabo5040536

Clauss, M. J., & Aarssen, L. W. (1994). Phenotypic Placity of Size--Fecundity Relationships in Arabidopsis Thaliana. *Journal of Ecology*, *82*(3), 447–455.

Cramer, G. R., Urano, K., Delrot, S., Pezzotti, M., & Shinozaki, K. (2011). Effects of abiotic stress on plants: A systems biology perspective. *BMC Plant Biology*, *11*. https://doi.org/10.1186/1471-2229-11-163

Cuevas, D. A., Garza, D., Sanchez, S. E., Rostron, J., Henry, C. S., Vonstein, V., … Edwards, R. A. (2019). Elucidating genomic gaps using phenotypic profiles [version 2 ; peer review : 1 approved , 1 approved with reservations], (May), 1–28.

Dadachova, E., Bryan, R. A., Huang, X., Moadel, T., Schweitzer, A. D., Aisen, P., … Casadevall, A. (2007). Ionizing radiation changes the electronic properties of melanin and enhances the growth of melanized fungi. *PLoS ONE*, *2*(5). https://doi.org/10.1371/journal.pone.0000457

Dasika, M. S., & Maranas, C. D. (2008). OptCircuit: An optimization based method for computational design of genetic circuits. *BMC Systems Biology*, *2*, 1–19. https://doi.org/10.1186/1752-0509-2-24

David, H., Özçelik, I. Ş., Hofmann, G., & Nielsen, J. (2008). Analysis of Aspergillus nidulans metabolism at the genome-scale. *BMC Genomics*, *9*, 1–15. https://doi.org/10.1186/1471-2164-9-163

Davidsohn, N., Beal, J., Kiani, S., Adler, A., Yaman, F., Li, Y., … Weiss, R. (2015). Accurate Predictions of Genetic Circuit Behavior from Part Characterization and Modular Composition. *ACS Synthetic Biology*, *4*(6), 673–681. https://doi.org/10.1021/sb500263b

de Felippes, F. F., McHale, M., Doran, R. L., Roden, S., Eamens, A. L., Finnegan, E. J., & Waterhouse, P. M. (2020). The key role of terminators on the expression and post-transcriptional gene silencing of transgenes. *Plant Journal*, *104*(1), 96–112. https://doi.org/10.1111/tpj.14907

De Martino, D., Capuani, F., Mori, M., De Martino, A., & Marinari, E. (2013). Counting and correcting thermodynamically infeasible flux cycles in genome-scale metabolic networks. *Metabolites*, *3*(4), 946–966. https://doi.org/10.3390/metabo3040946

de Oliveira Dal'Molin, C. G., Quek, L.-E., Palfreyman, R. W., Brumbley, S. M., & Nielsen, L. K. (2010). C4GEM, a Genome-Scale Metabolic Model to Study C4 Plant Metabolism. *Plant Physiology*, *154*(4), 1871–1885. https://doi.org/10.1104/pp.110.166488

Eisenman, H. C., & Casadevall, A. (2012). Synthesis and assembly of fungal melanin. *Applied Microbiology and Biotechnology*, *93*(3), 931–940. https://doi.org/10.1007/s00253-011-3777-2

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., … Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research*, *47*(D1), D427–D432. https://doi.org/10.1093/nar/gky995

English, M. A., Gayet, R. V., & Collins, J. J. (2021). Designing Biological Circuits: Synthetic Biology Within the Operon Model and Beyond. *Annual Review of Biochemistry*, *90*(1), 1–24. https://doi.org/10.1146/annurev-biochem-013118-111914

Exophiala dermatitidis (strain ATCC34100/CBS 525.76/NIH/UT8656). (2018).

Exophiala dermatitidis NIH/UT8656 Genome Assembly. (2011).

Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., … Palsson, B. (2007). A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Molecular Systems Biology*, *3*(121), 1–18. https://doi.org/10.1038/msb4100155

Feist, A. M., & Palsson, B. (2008). The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. *Nature Biotechnology*, *26*(6), 659–667. https://doi.org/10.1038/nbt1401

Fritzemeier, C. J., Hartleb, D., Szappanos, B., Papp, B., & Lercher, M. J. (2017). Erroneous energy-generating cycles in published genome scale metabolic networks: Identification and removal. *PLoS Computational Biology*, *13*(4), 1–14. https://doi.org/10.1371/journal.pcbi.1005494

Furumura, M., Solano, F., Matsunaga, N., Sakai, C., Spritz, R. A., & Hearing, V. J. (1998). Metal Ligand-Binding Specificities of the Tyrosinase-Related Proteins, *585*(242), 579–585.

García-Borrón, J. C., & Solano, F. (2002). Molecular anatomy of tyrosinase and its related proteins: Beyond the histidine-bound metal catalytic center. *Pigment Cell Research*, *15*(3), 162–173. https://doi.org/10.1034/j.1600-0749.2002.02012.x

Geis, P A, & Szaniszlo, P. J. (1984). Carotenoid pigments of the dematiaceous fungus Wangiella dermatitidis. *Mycologia*, *76*(2), 268–273.

Geis, Philip Anthony. (1981). Chemical composition of the yeast and sclerotic cell walls of Wangiella dermatitidis. *University of Texas at Austin*.

Gianchandani, E. P., Chavali, A. K., & Papin, J. A. (2010). The application of flux balance analysis in systems biology. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, *2*(3), 372–382. https://doi.org/10.1002/wsbm.60

Goldstein, G., Andrade, J. L., Meinzer, F. C., Holbrook, N. M., Cavelier, J., Jackson, P., & Celis, A. (1998). Stem water storage and diurnal patterns of water use in tropical forest canopy trees. *Plant, Cell and Environment*, *21*(4), 397–406. https://doi.org/10.1046/j.1365-3040.1998.00273.x

Gomes de Oliveira Dal'Molin, C., Quek, L.-E., Palfreyman, R. W., Brumbley, S. M., & Nielsen, L. K. (2010). AraGEM, a Genome-Scale Reconstruction of the Primary Metabolic Network in Arabidopsis. *Plant Physiology*, *152*, 579–589.

Gomes de Oliveira Dal'Molin, C., Quek, L.-E., Saa, P. A., & Nielsen, L. K. (2015). A multi-tissue genome-scale metabolic modeling framework for the analysis of whole plant systems. *Frontiers in Plant Science*, *6*(January), 1–12. https://doi.org/10.3389/fpls.2015.00004

Gonzali, S., Mazzucato, A., & Perata, P. (2009). Purple as a tomato: towards high anthocyanin tomatoes. *Trends in Plant Science*, *14*(5), 237–241. https://doi.org/10.1016/j.tplants.2009.02.001

Grafahrend-Belau, E., Junker, A., Eschenroder, A., Muller, J., Schreiber, F., & Junker, B. H. (2013). Multiscale Metabolic Modeling: Dynamic Flux Balance Analysis on a Whole-Plant Scale. *Plant Physiology*. https://doi.org/10.1104/pp.113.224006

Grafahrend-Belau, E., Schreiber, F., Koschutzki, D., & Junker, B. H. (2009). Flux Balance Analysis of Barley Seeds: A Computational Approach to Study Systemic Properties of Central Metabolism. *Plant Physiology*, *149*(1), 585–598. https://doi.org/10.1104/pp.108.129635

Gu, C., Kim, G. B., Kim, W. J., Kim, H. U., & Lee, S. Y. (2019). Current status and applications of genome-scale metabolic models. *Genome Biology*, *20*(1), 1–18. https://doi.org/10.1186/s13059-019-1730-3

Gudmundsson, S., Agudo, L., & Nogales, J. (2017). *Applications of genome-scale metabolic models of microalgae and cyanobacteria in biotechnology*. *Microalgae-Based Biofuels and Bioproducts: From Feedstock Cultivation to End-Products*. Elsevier Ltd. https://doi.org/10.1016/B978-0-08-101023-5.00004-2

Gudmundsson, Steinn, & Thiele, I. (2010). Computationally efficient flux variability analysis. *BMC Bioinformatics*, *11*(2), 2–4. https://doi.org/10.1186/1471-2105-11-489

Hall, R. D., Brouwer, I. D., & Fitzgerald, M. A. (2008). Plant metabolomics and its potential application for human nutrition. *Physiologia Plantarum*, *132*(2), 162–175. https://doi.org/10.1111/j.1399-3054.2007.00989.x

Hendrik Poorte, A., & Nagel, O. (2000). The role of biomass allocation in the growth response of plants to different levels of light, CO2, nutrients and water: A quantitative review. *Australian Journal of Plant Physiology*, *27*(189), 595–607. https://doi.org/10.1071/PP99173

Henry, C. S., Dejongh, M., Best, A. A., Frybarger, P. M., Linsay, B., & Stevens, R. L. (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature Biotechnology*, *28*(9), 977–982. https://doi.org/10.1038/nbt.1672

Herrgård, M. J., Fong, S. S., & Palsson, B. (2006). Identification of genome-scale metabolic network models using experimentally measured flux profiles. *PLoS Computational Biology*, *2*(7), 0676–0686. https://doi.org/10.1371/journal.pcbi.0020072

Hilder, V. A., & Boulter, D. (1999). Genetic engineering of crop plants for insect resistance - A critical review. *Crop Protection*, *18*(3), 177–191. https://doi.org/10.1016/S0261-2194(99)00028-9

Hill, A. D., Tomshine, J. R., Weeding, E. M. B., Sotiropoulos, V., & Kaznessis, Y. N. (2008). SynBioSS: The synthetic biology modeling suite. *Bioinformatics*, *24*(21), 2551–2553. https://doi.org/10.1093/bioinformatics/btn468

Holland, C. K., & Jez, J. M. (2018a). Arabidopsis: the original plant chassis organism. *Plant Cell Reports*, *37*(10), 1359–1366. https://doi.org/10.1007/s00299-018-2286-5

Holland, C. K., & Jez, J. M. (2018b). Arabidopsis: the original plant chassis organism. *Plant Cell Reports*, *37*(10), 1359–1366. https://doi.org/10.1007/s00299-018-2286-5

Hudjashov, G., Villems, R., & Kivisild, T. (2013). Global patterns of diversity and selection in human tyrosinase gene. *PloS One*, *8*(9), 1–14. https://doi.org/10.1371/journal.pone.0074307

Islam, M. M., Al-Siyabi, A., Saha, R., & Obata, T. (2018). Dissecting metabolic flux in C 4 plants: experimental and theoretical approaches. *Phytochemistry Reviews*, *17*(6), 1253–1274. https://doi.org/10.1007/s11101-018-9579-8

Ito, S., & Wakamatsu, K. (2011). Diversity of human hair pigmentation as studied by chemical analysis of eumelanin and pheomelanin. *Journal of the European Academy of Dermatology and Venereology*, *25*(12), 1369–1380. https://doi.org/10.1111/j.1468-3083.2011.04278.x

Ito, Shosuke. (2003). IFPCS presidential lecture: A chemist's view of melanogenesis. *Pigment Cell Research*, *16*(3), 230–236. https://doi.org/10.1034/j.1600-0749.2003.00037.x

Jacob, J. M., Karthik, C., Saratale, R. G., Kumar, S. S., Prabakar, D., Kadirvelu, K., & Pugazhendhi, A. (2018). Biological approaches to tackle heavy metal pollution: A survey of literature. *Journal of Environmental Management*, *217*, 56–70. https://doi.org/10.1016/j.jenvman.2018.03.077

Jeffery S. Amthor. (1984). The role of maintenance respiration in plant growth. *Plant, Cell and Environment*, *7*(8), 561–569.

Johnson, J. M.-F., Barbour, N. W., & Weyers, S. L. (2007). Chemical Composition of Crop Biomass Impacts Its Decomposition. *Soil Science Society of America Journal*, *71*(1), 155. https://doi.org/10.2136/sssaj2005.0419

Juenger, T. E., Mckay, J. K., Hausmann, N., Keurentjes, J. J. B., Sen, S., Stowe, K. A., … Richards, J. H. (2005). Identification and characterization of QTL underlying wholeplant physiology in Arabidopsis thaliana: δ13C, stomatal conductance and transpiration efficiency. *Plant, Cell and Environment*, *28*(6), 697–708. https://doi.org/10.1111/j.1365-3040.2004.01313.x

Kamaraj, B., & Purohit, R. (2014). Mutational analysis of oculocutaneous albinism: A compact review. *BioMed Research International*, *2014*. https://doi.org/10.1155/2014/905472

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017). KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, *45*(D1), D353–D361. https://doi.org/10.1093/nar/gkw1092

Karp, P. D., Weaver, D., & Latendresse, M. (2018). How accurate is automated gap filling of metabolic models? *BMC Systems Biology*, *12*(1), 1–11. https://doi.org/10.1186/s12918-018-0593-7

Khalil, A. S., & Collins, J. J. (2010). Synthetic biology: Applications come of age. *Nature Reviews Genetics*, *11*(5), 367–379. https://doi.org/10.1038/nrg2775

Khodayari, A., Chowdhury, A., & Maranas, C. D. (2015). Succinate Overproduction: A Case Study of Computational Strain Design Using a Comprehensive Escherichia coli Kinetic Model. *Frontiers in Bioengineering and Biotechnology*, *2*(January). https://doi.org/10.3389/fbioe.2014.00076

Kim, J., & Winfree, E. (2011). Synthetic in vitro transcriptional oscillators. *Molecular Systems Biology*, *7*(465), 1–15. https://doi.org/10.1038/msb.2010.119

Kim, T. Y., Sohn, S. B., Kim, Y. Bin, Kim, W. J., & Lee, S. Y. (2012). Recent advances in reconstruction and applications of genome-scale metabolic models. *Current Opinion in Biotechnology*, *23*(4), 617–623. https://doi.org/10.1016/j.copbio.2011.10.007

King, Z. A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J. A., … Lewis, N. E. (2016). BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Research*, *44*(D1), D515–D522. https://doi.org/10.1093/nar/gkv1049

Kumar, A. K., & Vatsyayan, P. (2010). Production of Lipid and Fatty Acids during Growth of Aspergillus terreus on Hydrocarbon Substrates, 1293–1300. https://doi.org/10.1007/s12010-009-8669-x

Kumar, J. (2018). Adaptations of Exophiala dermatitidis in stressful environments. *Dissertation, University of Nebraska - Lincoln*.

Kumar, J., Guo, P., Liu, K., Szaniszlo, P., & Harris, S. D. (2018). Role of Cdc42/Rac GTPase module in regualtion of the carotenoid biosynthetic gene cluster in the extremotolerant fungus Exophiala dermatitidis, 1–29.

Latendresse, M., & Karp, P. D. (2018). Evaluation of reaction gap-filling accuracy by randomization. *BMC Bioinformatics*, *19*(1), 1–13. https://doi.org/10.1186/s12859-018-2050-4

Leymarie, J., Lasceve, G., & Vavasseur, A. (1998). Interaction of stomatal responses to ABA and CO2 in Arabidopsis thaliana. *Australian Journal of Agricultural Research*, *49*, 317–327. https://doi.org/10.17700/jai.2015.6.1

Li, B., Suzuki, J.-I., & Hara, T. (1998). Latitudinal variation in plant size and relative growth rate in Arabidopsis thaliana. *Oecologia*, *115*, 293–301.

Limviphuvadh, V., Tan, C. S., Konishi, F., Jenjaroenpun, P., Xiang, J. S., Kremenska, Y., … Yong, W. P. (2018). Discovering novel SNPs that are correlated with patient outcome in a Singaporean cancer patient cohort treated with gemcitabine-based chemotherapy. *BMC Cancer*, *18*(1), 1–16. https://doi.org/10.1186/s12885-018-4471-x

Lin, P. C., Saha, R., Zhang, F., & Pakrasi, H. B. (2017). Metabolic engineering of the pentose phosphate pathway for enhanced limonene production in the cyanobacterium Synechocysti s sp. PCC 6803. *Scientific Reports*, *7*(1), 1–10. https://doi.org/10.1038/s41598-017-17831-y

Lipke, P. N., & Ovalle, R. (1998). Cell Wall Architecture in Yeast: New Structure and New Challenges. *Journal of Cateriology*, *180*(15), 3735–3740.

Liu, J., Gao, Q., Xu, N., & Liu, L. (2013). Genome-scale reconstruction and in silico analysis of Aspergillus terreus metabolism. *Molecular BioSystems*, *9*(7), 1939. https://doi.org/10.1039/c3mb70090a

Liu, W., & Stewart, C. N. (2015). Plant synthetic biology. *Trends in Plant Science*, *20*(5), 309–317. https://doi.org/10.1016/j.tplants.2015.02.004

Lonien, J., & Schwender, J. (2009). Analysis of metabolic flux phenotypes for two Arabidopsis mutants with severe impairment in seed storage lipid synthesis. *Plant Physiology*, *151*(3), 1617–1634. https://doi.org/10.1104/pp.109.144121

Luo, R. Y., Liao, S., Tao, G. Y., Li, Y. Y., Zeng, S., Li, Y. X., & Luo, Q. (2006). Dynamic analysis of optimality in myocardial energy metabolism under normal and ischemic conditions. *Molecular Systems Biology*, *2*, 1–6. https://doi.org/10.1038/msb4100071

Magnúsdóttir, S., Heinken, A., Kutt, L., Ravcheev, D. A., Bauer, E., Noronha, A., … Thiele, I. (2016). Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nature Biotechnology*, *35*(1), 81–89. https://doi.org/10.1038/nbt.3703

Mahadevan, R., Edwards, J. S., & Doyle, F. J. (2002). Dynamic Flux Balance Analysis of diauxic growth in Escherichia coli. *Biophysical Journal*, *83*(3), 1331–1340.

https://doi.org/10.1016/S0006-3495(02)73903-9

Mahmood, H. M., Mohammed, A. K., & Flayyih, M. T. A. (2015). Purification and Physiochemical Characterization of Pyomelanin Pigment Produced From Local Pseudomonas aeruginosa isolates. *World Journal of Pharmaceutical Research*, *4*(10), 289–299.

Maranas, C. D., & Zomorrodi, A. R. (2016). *Optimization Methods in Metabolic Networks*. Hoboken: Wiley.

Messing, J. (1998). Plant science in lac: A continuation of using tools from Escherichia coli in studying gene function in heterologous systems. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(1), 93–94. https://doi.org/10.1073/pnas.95.1.93

Min Lee, J., Gianchandani, E. P., Eddy, J. A., & Papin, J. A. (2008). Dynamic analysis of integrated signaling, metabolic, and regulatory networks. *PLoS Computational Biology*, *4*(5). https://doi.org/10.1371/journal.pcbi.1000086

Mintz-Oron, S., Meir, S., Malitsky, S., Ruppin, E., Aharoni, A., & Shlomi, T. (2012). Reconstruction of Arabidopsis metabolic network models accounting for subcellular compartmentalization and tissue-specificity. *Proceedings of the National Academy of Sciences*, *109*(1), 339–344. https://doi.org/10.1073/pnas.1100358109

Moreno, L. F., Vicente, V. A., & de Hoog, S. (2018). Black yeasts in the omics era: Achievements and challenges. *Medical Mycology*, *56*, 32–41. https://doi.org/10.1093/mmy/myx129

Mortimer, J. C. (2019). Plant synthetic biology could drive a revolution in biofuels and medicine. *Experimental Biology and Medicine*, *244*(4), 323–331. https://doi.org/10.1177/1535370218793890

Nagaya, S., Kawamura, K., Shinmyo, A., & Kato, K. (2010). The HSP terminator of arabidopsis thaliana increases gene expression in plant cells. *Plant and Cell Physiology*, *51*(2), 328–332. https://doi.org/10.1093/pcp/pcp188

National Center for Biotechnology Infornation. (n.d.). Retrieved from www.ncbi.nlm.nih.gov

Ng, C. Y., Jung, M., Lee, J., & Oh, M.-K. (2012). Production of 2,3-butanediol in Saccharomyces cerevisiae by in silico aided metabolic engineering. *Microbial Cell Factories*, *11*(1), 68. https://doi.org/10.1186/1475-2859-11-68

Nigam, R., & Liang, S. (2007). Algorithm for perturbing thermodynamically infeasible metabolic networks. *Computers in Biology and Medicine*, *37*(2), 126–133. https://doi.org/10.1016/j.compbiomed.2006.01.002

Nosanchuk, J. D., & Casadevall, A. (2006). Impact of Melanin on Microbial Virulence and Clinical Resistance to Antimicrobial Compounds. *Antimicrobial Agents and Chemotherapy*, *50*(11), 3519–3528.

Oakenfull, R. J., & Davis, S. J. (2017). Shining a light on the Arabidopsis circadian clock. *Plant, Cell & Environment*, *40*(11), 2571–2585. https://doi.org/10.1111/pce.13033

Oberhardt, M. A., Palsson, B., & Papin, J. A. (2009). Applications of genome-scOberhardt, M. A., Palsson, B., & Papin, J. A. (2009). Applications of genome-scale metabolic reconstructions. Molecular Systems Biology, 5(320), 1–15. https://doi.org/10.1038/msb.2009.77ale metabolic reconstructions. *Molecular Systems Biology*, *5*(320), 1–15. https://doi.org/10.1038/msb.2009.77

Ohkusu, M., Yamaguchi, M., Hata, K., Yoshida, S., Tanaka, R., Nishimura, K., … Takeo, K. (1999). Cellular and nuclear characteristics of Exophiala dermatitidis. *Studies in Mycology*, (43), 143–150.

Orth, J. D., Thiele, I., & Palsson, B. O. (2010). What is flux balance analysis? *Nature Publishing Group*, *28*(3), 245–248. https://doi.org/10.1038/nbt.1614

Ou-Yang, H., Stamatas, G., & Kollias, N. (2004). Spectral Responses of Melanin to Ultraviolet a Irradiation. *Journal of Investigative Dermatology*, *122*(2), 492–496. https://doi.org/10.1046/j.0022-202X.2004.22247.x

Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H. Y., Cohoon, M., … Vonstein, V. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research*, *33*(17), 5691–5702.

https://doi.org/10.1093/nar/gki866

Paolo, W. F., Dadachova, E., Mandal, P., Casadevall, A., Szaniszlo, P. J., & Nosanchuk, J. D. (2006). Effects of disrupting the polyketide synthase gene WdPKS1 in Wangiella [Exophiala] dermatitidis on melanin production and resistance to killing by antifungal compounds, enzymatic degredation, and extreme temperature. *BMC Microbiology*, *6*(55), 1–16.

Papadopoulos, J. S., & Agarwala, R. (2007). COBALT: Constraint-based alignment tool for multiple protein sequences. *Bioinformatics*, *23*(9), 1073–1079. https://doi.org/10.1093/bioinformatics/btm076

Pilalis, E., Chatziioannou, A., Thomasset, B., & Kolisis, F. (2011). An in silico compartmentalized metabolic model of Brassica napus enables the systemic study of regulatory aspects of plant central metabolism. *Biotechnology and Bioengineering*, *108*(7), 1673–1682. https://doi.org/10.1002/bit.23107

Pitkänen, E., Jouhten, P., Hou, J., Syed, M. F., Blomberg, P., Kludas, J., … Arvas, M. (2014). Comparative Genome-Scale Reconstruction of Gapless Metabolic Networks for Present and Ancestral Species. *PLoS Computational Biology*, *10*(2). https://doi.org/10.1371/journal.pcbi.1003465

Pixley, K. V., Falck-Zepeda, J. B., Giller, K. E., Glenna, L. L., Gould, F., Mallory-Smith, C. A., … Stewart, C. N. (2019). Genome editing, gene drives, and synthetic biology: Will they contribute to disease-resistant crops, and who will benefit? *Annual Review of Phytopathology*, *57*, 165–188. https://doi.org/10.1146/annurev-phyto-080417-045954

Poolman, M. G., Kundu, S., Shaw, R., & Fell, D. A. (2013). Responses to Light Intensity in a Genome-Scale Model of Rice Metabolism. *Plant Physiology*, *162*(2), 1060–1072. https://doi.org/10.1104/pp.113.216762

Poolman, Mark G, Miguet, L., Sweetlove, L. J., & Fell, D. A. (2009). A genome-scale metabolic model of Arabidopsis and some of its properties. *Plant Physiology*. https://doi.org/10.1104/pp.109.141267

Poyntner, C., Blasi, B., Arcalis, E., Mirastschijski, U., Sterflinger, K., & Tafer, H. (2016). The transcriptome of Exophiala dermatitidis during ex-vivo skin model infection. *Frontiers in Cellular and Infection Microbiology*, *6*. https://doi.org/10.1093/nar/gkn000

Price, N. D., Reed, J. L., & Palsson, B. (2004). Genome-scale models of microbial cells: Evaluating the consequences of constraints. *Nature Reviews Microbiology*, *2*(11), 886–897. https://doi.org/10.1038/nrmicro1023

Ranganathan, S., Suthers, P. F., & Maranas, C. D. (2010). OptForce: An optimization procedure for identifying all genetic manipulations leading to targeted overproductions. *PLoS Computational Biology*, *6*(4). https://doi.org/10.1371/journal.pcbi.1000744

Reed, J. L., & Palsson, B. (2003). Thirteen years of building constraint-based in silico models of Escherichia coli. *Journal of Bacteriology*, *185*(9), 2692–2699. https://doi.org/10.1128/JB.185.9.2692-2699.2003

Reed, J. L., Vo, T. D., Schilling, C. H., & Palsson, B. O. (2003). An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). *Genome Biology*, *4*(9), 1–12.

Regulation, C., & Major, M. (2007). Sugars and Circadian Regulation Make Major Contributions to the Global Regulation of Diurnal Gene Expression in Arabidopsis. *The Plant Cell*, *17*(December), 3257–3281. https://doi.org/10.1105/tpc.105.035261.1

Rizwan, M., Ali, S., Qayyum, M. F., Ok, Y. S., Zia-ur-Rehman, M., Abbas, Z., & Hannan, F. (2017). Use of Maize (Zea mays L.) for phytomanagement of Cd-contaminated soils: a critical review. *Environmental Geochemistry and Health*, *39*(2), 259–277. https://doi.org/10.1007/s10653-016-9826-0

Robinson, J. L., Kocabaş, P., Wang, H., Cholley, P. E., Cook, D., Nilsson, A., … Nielsen, J. (2020). An atlas of human metabolism. *Science Signaling*, *13*(624), 1–12. https://doi.org/10.1126/scisignal.aaz1482

Saa, P. A., & Nielsen, L. K. (2016). Fast-SNP: A fast matrix pre-processing algorithm for efficient loopless flux optimization of metabolic models. *Bioinformatics*, *32*(24), 3807–3814.

https://doi.org/10.1093/bioinformatics/btw555

Saha, R., Liu, D., Connor, A. H., Liberton, M., Yu, J., & Bhattacharyya-pakrasi, M. (2016). Diurnal Regulation of Cellular Processes in the Cyanobacterium Synechocystis sp . Strain PCC 6803 : Insights from Transcriptomic ,. *MBio*, *7*(3), 1–14. https://doi.org/10.1128/mBio.00464-16.Editor

Saha, R., Suthers, P. F., & Maranas, C. D. (2011). Zea mays irs1563: A comprehensive genome-scale metabolic reconstruction of maize metabolism. *PLoS ONE*, *6*(7). https://doi.org/10.1371/journal.pone.0021784

Saha, R., Verseput, A. T., Berla, B. M., Mueller, T. J., Pakrasi, H. B., & Maranas, C. D. (2012). Reconstruction and Comparison of the Metabolic Potential of Cyanobacteria Cyanothece sp. ATCC 51142 and Synechocystis sp. PCC 6803. *PLoS ONE*, *7*(10). https://doi.org/10.1371/journal.pone.0048285

Santiago, J. P., & Tegeder, M. (2016). Connecting source with sink: The role of arabidopsis AAP8 in phloem loading of amino acids. *Plant Physiology*, *171*(1), 508–521. https://doi.org/10.1104/pp.16.00244

Satish Kumar, V., Dasika, M. S., & Maranas, C. D. (2007). Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics*, *8*, 1–16. https://doi.org/10.1186/1471-2105-8-212

Schellenberger, J., Lewis, N. E., & Palsson, B. (2011). Elimination of thermodynamically infeasible loops in steady-state metabolic models. *Biophysical Journal*, *100*(3), 544–553. https://doi.org/10.1016/j.bpj.2010.12.3707

Scheller, L., Schmollack, M., Bertschi, A., Mansouri, M., Saxena, P., & Fussenegger, M. (2020). Phosphoregulated orthogonal signal transduction in mammalian cells. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-16895-1

Schmaler-Ripcke, J., Sugareva, V., Gebhardt, P., Winkler, R., Kniemeyer, O., Heinekamp, T., & Brakhage, A. A. (2009). Production of pyomelanin, a second type of melanin, via the tyrosine degradation pathway in Aspergillus fumigatus. *Applied and Environmental Microbiology*, *75*(2), 493–503. https://doi.org/10.1128/AEM.02077-08

Schnitzler, N., Peltroche-Llacsahuanga, H., Bestier, N., Zundorf, J., Lutticken, R., & Haase, G. (1999). Effect of melanin and carotenoids of *Exophiala* (*Wangiella*) *dermatitidis* on phagocytosis, oxidative burst, and killing by human neutrophils. *Infection and Immunity*, *67*(1), 94–101.

Schoch, C. L., Sung, G.-H., Lopez-Giraldez, F., Townsend, J. P., Miadlikowska, J., Hofstetter, V., & Gueidan, C. (2009). The ascomycota tree of life: a phylum-wide phylogeny clarifies the origin and evolution of fundamental reproductive and ecological traits. *Systems Biology*, *58*.

Schroeder, W. L., & Saha, R. (2020a). OptFill: A Tool for Infeasible Cycle-Free Gapfilling of Stoichiometric Metabolic Models. *IScience*, *23*(1), 100783. https://doi.org/10.1016/j.isci.2019.100783

Schroeder, W. L., & Saha, R. (2020b). OptFill: A Tool for Infeasible Cycle-Free Gapfilling of Stoichiometric Metabolic Models. *IScience*, *23*(1), 100783. https://doi.org/10.1016/j.isci.2019.100783

Schulze, E. (1986). Vapor Exchange in Response To Drought in the Atmosphere and in the Soil 1. *Plant Physiol*, *37*, 247–274.

Sekara, A., Poniedziałek, M., Ciura, J., & Jedrszczyk, E. (2005). Zinc and copper accumulation and distribution in the tissues of nine crops: Implications for phytoremediation. *Polish Journal of Environmental Studies*, *14*(6), 829–835.

Sengupta, S., & Majumder, A. L. (2014). Physiological and genomic basis of mechanical-functional trade-off in plant vasculature. *Frontiers in Plant Science*, *5*(MAY). https://doi.org/10.3389/fpls.2014.00224

Shaw, R., & Cheung, C. Y. M. (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*(June), 1–15. https://doi.org/10.3389/fpls.2018.00884

Shipley, B., & Vu, T.-T. (2002). Dry matter content as a measure of dry matter concentration in plants and their parts. *New Phytologist*, *153*, 359–364. https://doi.org/10.15448/1984-4301.2017.2.26404

Shoaie, S., Karlsson, F., Mardinoglu, A., Nookaew, I., Bordel, S., & Nielsen, J. (2013). Understanding the interactions between bacteria in the human gut through metabolic modeling. *Scientific Reports*, *3*(2532), 1–10. https://doi.org/10.1038/srep02532

Simons, M., Saha, R., Amiour, N., Kumar, A., Guillard, L., Clement, G., … Maranas, C. D. (2014). Assessing the Metabolic Impact of Nitrogen Availability Using a Compartmentalized Maize Leaf Genome-Scale Model. *Plant Physiology*, *166*(3), 1659–1674. https://doi.org/10.1104/pp.114.245787

Singh, S., Malhotra, A. G., Pandey, A., & Pandey, K. M. (2013). Computational model for pathway reconstruction to unravel the evolutionary significance of melanin synthesis. *Bioinformation*, *9*(2), 94–100. https://doi.org/10.6026/97320630009094

Solano, F. (2014). Melanins: Skin Pigments and Much More—Types, Structural Models, Biological Functions, and Formation Routes. *New Journal of Science*, *2014*, 1–28. https://doi.org/10.1155/2014/498276

Solovchenko, A. E., & Merzlyak, M. N. (2008). Screening of visible and UV radiation as a photoprotective mechanism in plants. *Russian Journal of Plant Physiology*, *55*(6), 719–737. https://doi.org/10.1134/s1021443708060010

Spritz, R. A. (1994). Molecular genetics of oculocutaneous albinism. *Human Molecular Genetics*, *3*(9), 1469–1475.

Spritz, R. A., Ho, L., Furumura, M., & Hearing, V. J. (1997). Mutational analysis of copper binding by human tyrosinase. *Journal of Investigative Dermatology*, *109*(2), 207–212. https://doi.org/10.1111/1523-1747.ep12319351

Srinivasan, S., Cluett, W. R., & Mahadevan, R. (2015). Constructing kinetic models of metabolism at genome-scales: A review. *Biotechnology Journal*, *1359*, 1345–1359.

https://doi.org/10.1002/biot.201400522

Stolyar, S., Van Dien, S., Hillesland, K. L., Pinel, N., Lie, T. J., Leigh, J. A., & Stahl, D. A. (2007). Metabolic modeling of a mutualistic microbial community. *Molecular Systems Biology*, *3*(92), 1–14. https://doi.org/10.1038/msb4100131

Strobel, I., Breitenbach, J., Scheckhuber, C. Q., Osiewacz, H. D., & Sandmann, G. (2009). Carotenoids and carotenogenic genes in Podospora anserina: Engineering of the carotenoid composition extends the life span of the mycelium. *Current Genetics*, *55*(2), 175–184. https://doi.org/10.1007/s00294-009-0235-0

Sudhadham, M., Prakitsin, S., Sivichai, S., Chaiyarat, R., Dorrestein, G. M., Menken, S. B. J., & de Hoog, G. S. (2008). The neurotropic black yeast Exophiala dermatitidis has a possible origin in the tropical rain forest. *Studies in Mycology*, *61*, 145–155. https://doi.org/10.3114/sim.2008.61.15

Szaniszlo, P. J. (2002). Molecular genetic studies of the model dematiaceous pathogen Wangiella dermatitidis. *International Journal of Medical Microbiology : IJMM*, *292*(5–6), 381–390. https://doi.org/10.1078/1438-4221-00221

Tan, S. I., & Ng, I. S. (2021). CRISPRi-Mediated NIMPLY Logic Gate for Fine-Tuning the Whole-Cell Sensing toward Simple Urine Glucose Detection. *ACS Synthetic Biology*, *10*(2), 412–421. https://doi.org/10.1021/acssynbio.1c00014

Tegeder, M., & Hammes, U. Z. (2018). The way out and in: phloem loading and unloading of amino acids. *Current Opinion in Plant Biology*, *43*, 16–21. https://doi.org/10.1016/j.pbi.2017.12.002

Terzer, M., Maynard, N. D., & Covert, M. W. (2009). Genome-scale metabolic networks, (December). https://doi.org/10.1002/wsbm.037

Thiele, I., & Palsson, B. Ø. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols*, *5*(1), 93–121. https://doi.org/10.1038/nprot.2009.203

Thornley, J. H. M., & Cannell, M. G. R. (1999). Modelling the Components of Plant Respiration: Representation and Realism. *Annals of Botany*, *85*, 55–56.

Thornley, J. H. M., & Cannell, M. G. R. (2000). Managing forest for wood yield and carbon storage: A theoretical study. *Tree Physiology*, *20*(February), 477–484.

Toledo, A. V., Franco, M. E. E., Yanil Lopez, S. M., Troncozo, M. I., Saparrat, M. C. N., & Balatti, P. A. (2017). Melanins in fungi: Types, localization and putative biological roles. *Physiological and Molecular Plant Pathology*, *99*, 2–6. https://doi.org/10.1016/j.pmpp.2017.04.004

UniProtKB. (2018). E. coli K12. Retrieved August 20, 2008, from www.uniprot.org/uniprot/?query=E.+coli+K-12+strain+1655&sort=score

Upadhyay, S., Xu, X., Lowry, D., Jackson, J. C., Roberson, R. W., & Lin, X. (2016). Subcellular Compartmentalization and Trafficking of the Biosynthetic Machinery for Fungal Melanin. *Cell Reports*, *14*(11), 2511–2518. https://doi.org/10.1016/j.celrep.2016.02.059

Vardhan, K. H., Kumar, P. S., & Panda, R. C. (2019a). A review on heavy metal pollution, toxicity and remedial measures: Current trends and future perspectives. *Journal of Molecular Liquids*, *290*, 111197. https://doi.org/10.1016/j.molliq.2019.111197

Vardhan, K. H., Kumar, P. S., & Panda, R. C. (2019b). A review on heavy metal pollution, toxicity and remedial measures: Current trends and future perspectives. *Journal of Molecular Liquids*, *290*, 111197. https://doi.org/10.1016/j.molliq.2019.111197

Vareda, J. P., Valente, A. J. M., & Durães, L. (2019). Assessment of heavy metal pollution from anthropogenic activities and remediation strategies: A review. *Journal of Environmental Management*, *246*(June), 101–118. https://doi.org/10.1016/j.jenvman.2019.05.126

Voigt, C. A. (2020). Synthetic biology 2020–2030: six commercially-available products that are changing our world. *Nature Communications*, *11*(1), 10–15. https://doi.org/10.1038/s41467-020-20122-2

Vongsangnak, W., Olsen, P., Hansen, K., Krogsgaard, S., & Nielsen, J. (2008). Improved annotation through genome-scale metabolic modeling of Aspergillus oryzae. *BMC Genomics*, *9*, 1–14. https://doi.org/10.1186/1471-2164-9-245

Williams, T. C. R., Poolman, M. G., Howden, A. J. M., Schwarzlander, M., Fell, D. A., Ratcliffe, R. G., & Sweetlove, L. J. (2010). A Genome-Scale Metabolic Model Accurately Predicts Fluxes in Central Carbon Metabolism under Stress Conditions. *Plant Physiology*, *154*(1), 311–323. https://doi.org/10.1104/pp.110.158535

Wuana, R. A., & Okieimen, F. E. (2010). Phytoremediation Potential of Maize (Zea mays L.). A Review. *African Studies on Population and Health*, *00*(4), 275–287. Retrieved from http://www.asopah.org

Xia, P. F., Ling, H., Foo, J. L., & Chang, M. W. (2019). Synthetic genetic circuits for programmable biological functionalities. *Biotechnology Advances*, *37*(6), 107393. https://doi.org/10.1016/j.biotechadv.2019.04.015

Yamamoto, H. Y., & Bangham, A. D. (1978). Carotenoid organization in membranes. Thermal transition and spectral properties of carotenoid-containing liposomes. *BBA - Biomembranes*, *507*(1), 119–127. https://doi.org/10.1016/0005-2736(78)90379-6

Yizhak, K., Chaneton, B., Gottlieb, E., & Ruppin, E. (2015). Modeling cancer metabolism on a genome scale. *Molecular Systems Biology*, *11*(6), 817. https://doi.org/10.15252/msb.20145307

Yu, C., & Lin, C. (2004). Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n -peptide compositions, 1402–1406. https://doi.org/10.1110/ps.03479604.calization

Yu, C. S., Lin, C. J., & Hwang, J. K. (2006). Prediction of protein subcellular localization. *Proteins: Structure, Function and Bioinformatics*, *64*, 643–651.

Zaghdoudi, K., Ngomo, O., Vanderesse, R., Arnoux, P., Myrzakhmetov, B., Frochot, C., & Guiavarc'h, Y. (2017). Extraction, Identification and Photo-Physical Characterization of Persimmon (Diospyros kaki L.) Carotenoids. *Foods*, *6*(1), 4.

https://doi.org/10.3390/foods6010004

Zhang, L., Tan, Q., Lee, R., Trethewy, A., Lee, Y. H., & Tegeder, M. (2010). Altered xylem-phloem transfer of amino acids affects metabolism and leads to increased seed yield and oil content in Arabidopsis. *Plant Cell*, *22*(11), 3603–3620. https://doi.org/10.1105/tpc.110.073833

Zhang, X., Tervo, C. J., & Reed, J. L. (2016). Metabolic assessment of E. coli as a Biofactory for commercial products. *Metabolic Engineering*, *35*, 64–74. https://doi.org/10.1016/j.ymben.2016.01.007

Zomorrodi, A. R., & Maranas, C. D. (2012). OptCom: A multi-level optimization framework for the metabolic modeling and analysis of microbial communities. *PLoS Computational Biology*, *8*(2), 1–13. https://doi.org/10.1371/journal.pcbi.1002363

Zomorrodi, A. R., & Maranas, C. D. (2014a). Coarse-grained optimization-driven design and piecewise linear modeling of synthetic genetic circuits. *European Journal of Operational Research*, *237*(2), 665–676. https://doi.org/10.1016/j.ejor.2014.01.054

Zomorrodi, A. R., & Maranas, C. D. (2014b). Coarse-grained optimization-driven design and piecewise linear modeling of synthetic genetic circuits. *European Journal of Operational Research*, *237*(2), 665–676. https://doi.org/10.1016/j.ejor.2014.01.054