A COMPUTATIONAL AND EXPERIMENTAL INVESTIGATION OF LIGNIN METABOLISM IN ARABIDOPSIS

by

Rohit Jaini

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Davidson School of Chemical Engineering

West Lafayette, Indiana

December 2017

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. John A. Morgan, Chair
Davidson School of Chemical Engineering
Dr. Natalia Dudareva
Department of Biochemistry
Dr. Doraiswami Ramkrishna
Davidson School of Chemical Engineering
Dr. Rajamani Gounder
Davidson School of Chemical Engineering

Approved by:
Dr. John A. Morgan
Head of the Graduate Program

This thesis is dedicated to my parents;
the unsung heroes who have shaped my life,
Raghu and Neetha

ACKNOWLEDGMENTS

First and foremost, I would like to convey my deepest gratitude and respect to my advisor Dr. John A. Morgan. This thesis wouldn't have been possible without his invaluable guidance and mentorship. His approach to research, acute attention to detail, a striking grasp on a variety of fields in science, and an attitude that fosters collaborations, have helped me grow as a researcher and a scientist. He has always been welcoming of new ideas and has pushed me to give my best at every juncture. A fond memory I would always cherish is when we conceived and hashed out an idea for a proposal at 2:30 a.m in your office. I can't thank you enough for being a wonderful mentor and guide, and making this journey in grad school a memorable one.

A sincere thanks goes out to Dr. Natalia Dudareva and Dr. Clint Chapple for being my co-advisors. Both of you epitomize the passion and grit that goes into conducting breakthrough research. It has been an honor being mentored by such giants of science. Next, I would like to thank Dr. Doraiswami Ramkrishna and Dr. Rajamani Gounder for serving on my committee and for encouraging me to critically look into my research.

I would like to thank my fellow graduate students, Longyun Guo and Peng Wang, who have been on this journey with me from the start. In addition to being some of the smartest graduate students I've had the wonderful opportunity to work with, you both are, to me, examples of model researchers and have constantly inspired me to work harder at every stage. A huge thanks to fellow Morgan lab members, Robin Wheeler, Rick Ray, Jeremiah Vue, and Joel King for their inputs, support and for persevering through all my long presentations. I would like to thank Agnes Mendonca for being my coffee bud, my reading bud, my crutch, and above all for making the long hours in FRNY tolerable.

A huge shout out to all the undergrads I had the opportunity to mentor and work with. I would specifically like to thank Joseph Parry for helping me gather data and conduct crucial experiments towards my dissertation. I've been lucky to have had such an independent, hardworking, and sincere undergrad during the most critical time of my PhD.

Last, but most definitely not the least, I would like to thank my friends and family, specifically my parents, for supporting me at every stage in life and making me a better person.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Author: Jaini, Rohit. PhD
Institution: Purdue University
Degree Received: December 2017
Title: A Computational and Experimental Investigation of Lignin Metabolism in *Arabidopsis thaliana*.
Major Professor: John A. Morgan

Predominantly localized in plant secondary cell walls, lignin is a highly cross-linked, aromatic polymer that imparts structural support to plant vasculature, and renders biomass recalcitrant to pretreatment techniques impeding the economical production of biofuels. Lignin is synthesized via the phenylpropanoid pathway where the primary precursor phenylalanine (Phe) undergoes a series of functional modifications catalyzed by 11 enzyme families to produce *p*-coumaryl, coniferyl, and sinapyl alcohol, which undergo random polymerization into lignin. Several metabolic engineering efforts have aimed to alter lignin content and composition, and make biofuel feedstock more amenable to pretreatment techniques. Despite significant advances, several questions pertaining to carbon flux distribution in the phenylpropanoid network remain unanswered. Furthermore, complexity of the metabolic pathway and a lack of sensitive analytical tools add to the challenges of mechanistically understanding lignin synthesis.

In this work, I describe improvements in analytical techniques used to characterize phenylpropanoid metabolism that have been applied to obtain a comprehensive quantitative mass balance of the phenylpropanoid pathway. Finally, machine learning and artificial intelligence were utilized to make predictions about optimal lignin amount and composition for improving saccharification. In summary, the overarching goal of this thesis

was to further the understanding of lignin metabolism in the model system, *Arabidopis thaliana*, employing a combination of experimental and computational strategies.

First, we developed comprehensive and sensitive analytical methods based on liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) to quantify intermediates of the phenylpropanoid pathway. Compared to existing targeted profiling techniques, the methods were capable of quantifying a wider range of phenylpropanoid intermediates, at lower concentrations, with minimal sample preparation. The technique was used to generate flux maps for wild type and mutant Arabidopsis stems that were fed exogenously $^{13}C_6$-Phe. Flux maps computed in this work; (i) suggest the presence of a hitherto uncharacterized alternative route to caffeic acid and lignin synthesis, (ii) shed light on flux splits at key branch points of the network, and (iii) indicate presence of inactive pools for a number of metabolites.

Finally, we present a machine learning based model that captures the non-linear relationship between lignin content and composition, and saccharification efficiency. A support vector machine (SVM) based regression technique was developed to predict saccharification efficiency and biomass yields as a function of lignin content, and composition of monomers that make up lignin, namely *p*-coumaryl (H), coniferyl (G), and sinapyl (S) alcohol derived lignin. The model was trained on data obtained from the literature and validated on Arabidopsis mutants that were excluded from the training data set. Functional forms obtained from SVM regression were further optimized using genetic algorithms (GA) to maximize total sugar yields. Our efforts resulted in two optimal solutions with lower lignin content and interestingly varying H:G:S composition that were conducive to saccharide extractability.

1. INTRODUCTION

1.1 Background

1.1.1 Biofuels from lignocellulosic feedstock – lignin recalcitrance

One of the greatest challenges of the twenty first century is to provide clean and sustainable sources of fuels and chemicals to bridge the growing gap between energy consumption and dwindling fossil fuel reserves[1]. The skewed energy supply-demand balance is only exacerbated by uncertainty in petroleum supplies, and increasing green-house gas emissions and global warming concerns associated with the use of fossil fuels. As a result, there has been a cognizant shift towards the use of renewable and alternative sources of energy such as hydroelectric, solar, wind, and biomass. Biomass derived energy in the form of biofuels (encompassing bioethanol, bio-oil, and biodiesel), is unique amongst all the available alternative energy sources in its direct compatibility with existing liquid transportation fuel[2,3]. Bioethanol is predominantly produced by fermentation of depolymerized sugars from plant material. Second generation biofuels from lignocellulosic feedstocks have gained significant attention as they offer benefits such as (i) abundance of raw material, (ii) consumption of inedible parts and agricultural waste residues, and (iii) growth on abandoned and marginal lands[4,5].

Lignocellulose is primarily made up of lignin, cellulose, hemicellulose, pectin, and proteins. Biofuel production from lignocellulosic feedstock requires depolymerization of cellulosic sugars by enzyme hydrolysis – a process known as saccharification – for further conversion to ethanol[6,7]. Lignin, a hetero-aromatic polymer predominantly localized in plant secondary cell walls renders biofuel feedstock recalcitrant to microbial and enzymatic

digestion as it crosslinks with cellulose and hemi-cellulose essentially entrapping useful cell-wall polysaccharides limiting their accessibility to cellulases[7,8]. In order to achieve effective hydrolysis, feedstock is subjected to necessary pretreatment techniques to loosen lignin's 'grip' on cell wall polysaccharides in turn making biofuel production a cost intensive process. Pretreatment technologies span mechanical pretreatments, physicochemical pretreatments, chemical pretreatments, and biological pretreatments[9].

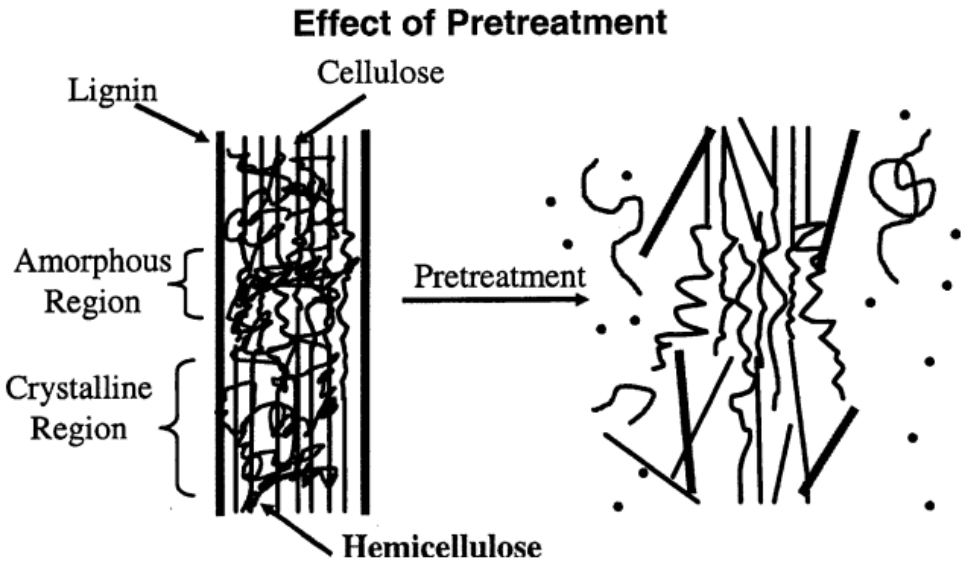


Figure 1.1: Goal of pretreatment techniques on lignocellulosic material.

Pretreatment costs alone account for almost 20% of the total cost of production of 1 gallon of ethanol[9,10]. Reducing these pretreatment costs would go a long way in making lignocellulose derived biofuels an economically viable alternative to existing fossil fuels. There has been ample research focused on optimizing pretreatment technologies, hydrolysis, and fermentation of biomass[11–13]. But with increasing understanding of lignin biosynthesis and other cell wall components, genetic modifications of plant cell wall to obtain feedstock with improved saccharification efficiency has been made

possible[8,14–18]. An introduction to lignin biosynthesis and a description of some noteworthy lignin engineering efforts have been summarized in the following sections.

1.1.2 Lignin biosynthesis in plants

Lignin is synthesized via the phenylpropanoid pathway – also referred to as the monolignol biosynthesis pathway – which owing to decades of biochemical and genetic investigations has been well characterized (Figure 1.2, [19]). In addition to the synthesis of monolignols: *p*-coumaryl alcohol, coniferyl alcohol, and sinapyl alcohol, the fundamental building blocks of lignin, many phenylpropanoid intermediates serve as precursors to other useful biochemical such as flavonoids, coumarins, tannins, hydroxycinnamic acid derivatives and lignans[20]. The pathway begins with the deamination of Phe to cinnamic acid by Phe ammonia lyase (PAL), a reaction that essentially bridges primary metabolism to phenylpropanoid metabolism and is considered a limiting step in directing carbon flux towards lignin[21]. Cinnamic acid hydroxylation at the *para* position by C4H followed by CoA thioester formation by 4-coumarate:CoA ligase 1 (4CL1) – one of the four isoforms identified in Arabidopsis – results in the formation so *p*-coumaroyl CoA[22]. Apart from being an essential precursor leading to a diverse set of secondary metabolites like flavonoids, stilbenes and tannins, *p*-coumaroyl CoA is also the first key branch point in lignin synthesis where carbon is routed to G and S subunit production (Figure 1.2).

Hydroxycinnamoyl CoA:shikimate hydroxycinnamoyl transferase (HCT) and *p*-coumaroylshikimate 3′-hydroxylase (C3′H) catalyze the set of bridging reactions between H, and G&S subunits. HCT is a reversible enzyme that catalyzes conversion of *p*-coumaroyl CoA to *p*-coumaryl shikimate and caffeoyl shikimate to caffeoyl CoA. Shikimic

acid is consumed in the former and released in the latter reaction[23,24]. It has been proposed that shikimic acid may be a putative regulatory link between phenylalanine synthesis (by the shikimate pathway) in the plastid and its utilization (by the phenylpropanoid pathway) in the cytosol. Hydrolysis of caffeoyl shikimate followed by CoA ligation provides an alternative route to the synthesis of caffeoyl CoA, that bypasses the second HCT reaction, making caffeoyl-shikimate the second key branch point in lignin synthesis[25].

Caffeoyl CoA subsequently undergoes methylation by caffeoyl CoA 3-O-methyltransferase (CCoAOMT) to produce feruloyl CoA, which is reduced to coniferaldehyde by cinnamoyl-CoA reductase (CCR), the third key branch point in lignin biosynthesis (Figure 1.2). Coniferaldehyde undergoes a series of reductions, hydroxylations and methylation reactions, catalyzed by cinnamyl alcohol dehydrogenase (CAD), ferulate 5-hydroxylase (F5H), caffeic acid/5-hydroxyferulic acid O-methyltransferase (COMT) respectively, to form coniferyl and sinapyl alcohol[19].

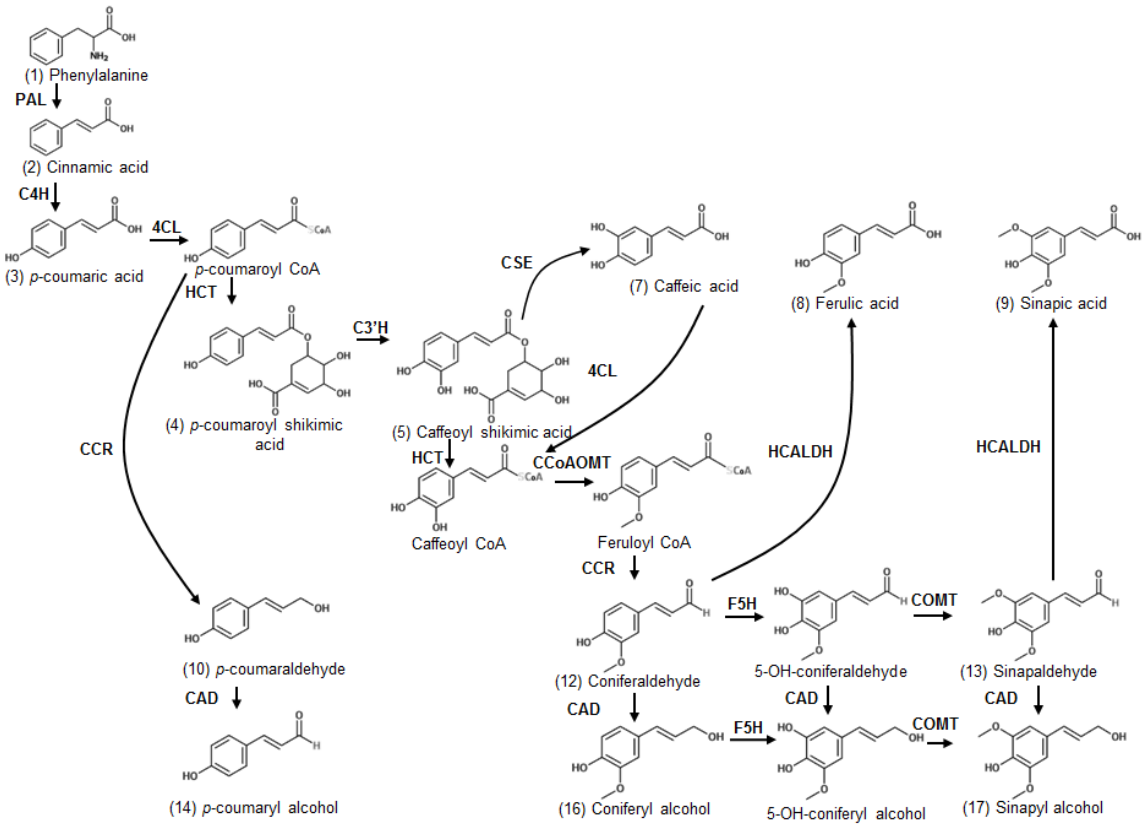


Figure 1.2: The phenylpropanoid pathway in Arabidopsis. PAL, phenylalanine ammonia-lyase; C4H, cinnamic acid 4-hydroxylase; 4CL, 4-coumarate:CoA ligase; HCT, hydroxycinnamoyl-coenzyme A shikimate:quinate hydroxycinnamoyl-transferase; C3'H, p-coumaroyl shikimate 3'-hydroxylase; CCoAOMT, caffeoyl CoA 3-O-methyltransferase; CCR, cinnamoyl-CoA reductase; F5H, ferulate 5-hydroxylase; COMT, caffeic acid/5-hydroxyferulic acid O-methyltransferase; CAD, cinnamyl alcohol dehydrogenase; HCALDH, hydroxycinnamaldehyde dehydrogenase. The reaction catalyzed by HCALDH leads to the synthesis of ferulic and sinapic acid.

1.1.3 Lignin engineering

In the past two decades, several metabolic engineering efforts have targeted lignin synthesis by manipulating the expression of individual genes in the phenylpropanoid pathway[15,16,26,27]. From a lignocentric point of view, consequences of the host of engineering efforts across different plant systems can be broadly categorized into one or more of the following categories; (i) reduced lignin, (ii) altered lignin composition, and

(iii) incorporation of unconventional metabolites into lignin. In general downregulation of PAL, C4H, 4CL, HCT, C3H, CSE, CCoAOMT, CCR, and to some extent, CAD, have a significant effect on lignin content[6]. Drastic changes in lignin composition can be engineering with plants deficient in C3′H, HCT, and CSE are strikingly enriched in H lignin, which constitutes the minor component in wild-type plants[25,28,29]. Downregulation and overexpression of F5H results in lignin essentially composed of G and S lignin respectively. Unconventional intermediates such as 5-OH-coniferyl alcohol are incorporated into lignin in COMT downregulated lines, while CAD downregulated plants have significant incorporation of hydroxycinnamaldehydes in lignin[30–32].

Most of the genetic engineering experiments previously discussed were predicated solely on the identity of a specific enzyme in the pathway, and its (at times assumed) role in lignin synthesis. The wide spectrum of phenotypes obtained as a result these experiments stems from a disparate and distributed flux control across enzymes of the network, inherent regulation in phenylpropanoid metabolism, and complex interactions with other metabolic networks, which still remains to be understood. Therefore, despite the commendable strides made in lignin manipulation, several questions regarding carbon flux control have been raised.

From the point of view of saccharification, in general a reduced lignin phenotype resulted in increased biomass digestibility, but no reasonable correlations between lignin composition and saccharification efficiencies were observed[33]. It was previously reported that a high S/G ratio is conducive for improved saccharification[14,33], but there have been reports of transgenic lines that exhibited lower or an unchanged saccharification phenotype despite having a high S/G ratio[16,32]. In addition, lines with high H lignin

units were characterized by higher saccharification efficiencies[25,28]. All previous studies that analyzed the relation between these biological traits and saccharification employing one to one correlations and linearly mapping the relation between the variables[33–35]. It is evident that a multi-variate approach is necessary in understanding the non-linear dependence of biomass digestibility on lignin content and composition.

1.2 Motivation and Research Objectives

The strikingly different phenotypes obtained as a result of downregulation of different enzymes of the phenylpropanoid pathway, clearly indicate a more distributed control of carbon flux to lignin. Furthermore, unanticipated pleiotropic effects, complexity of the metabolic pathway, lack of sensitive analytical tools to measure low metabolite concentrations, sub-cellular compartmentation of phenylpropanoid intermediates, and a rigid regulatory hierarchy add to the challenges of mechanistically understanding lignin synthesis. In addition, there is no thorough understanding of how a certain lignin phenotype contributes to the modification in biomass saccharifiability.

Although a combination of systems biology and integrative ‘omics’ approaches to address all the issues listed above and to gain a more mechanistic understanding of lignin biosynthesis; in this work, we present experimental and computational strategies to investigate and further the understanding of (i) flux control in lignin metabolism; and (ii) relation between lignin content and composition, and biomass digestibility in the model system, *Arabidopis thaliana*. Research conducted in this work is aimed at addressing the following objectives:

Objective 1: To develop analytical techniques for extensively and accurately quantifying intermediates of the phenylpropanoid pathway.

Objective 2: To estimate the relative sub-cellular distribution of metabolites in different plant cell organelles of Arabidopsis stems using non-aqueous fractionation (NAQF).

Objective 3: To compute high resolution flux maps of the metabolic network across different Arabidopsis genetic backgrounds, using ^{13}C-metabolic flux analysis (MFA).

Objective 4: To evaluate the functional forms that relate lignin content and composition to saccharification efficiency and growth, and further estimate an optimal lignin phenotype that maximizes the total saccharification yields using a combination of machine learning and evolutionary computation.

1.3 Organization of Dissertation

This dissertation is organized as follows. In chapters 2 and 3, details of analytical method development for quantifying phenylpropanoid pathway intermediates have been extensively covered. The methods presented in these chapters have been used to measure metabolite concentrations in all subsequent studies discussed in this work. In chapter 4, a modeling and experimental strategy to compute flux maps in Arabidopsis stems is presented using dynamic isotopic labeling measurement of phenylpropanoid metabolites after exogenously supplying $^{13}C_6$-Phe. Chapter 5 establishes the application of the analytical tools developed in this study on different genetic background of Arabidopsis. A comparative analysis of soluble metabolite pools across different genotypes has been conducted using wild-type Arabidopsis plants as a reference. In chapter 6, non-aqueous

fractionation technique (NAQF) and its application to Arabidopsis stems has been described. NAQF has been used to estimate the relative distribution of key metabolites of the phenylpropanoid pathway across different sub-cellular compartments. Following this, Chapter 7 delineates a mathematical modeling strategy that is a combination of machine learning and evolutionary computation has been presented to map the relationship between lignin content and composition, and saccharification efficiency. Optimal lignin content and composition were estimated that would maximize the total sugar yields. Chapter 8, is the final chapter in which a brief summary of future research directions and recommendations have been summarized.

2. ANALYTICAL METHOD DEVELOPMENT (I): QUANTIFYING INTERMEDIATES OF THE PHENYLPROPANOID PATHWAY.

2.1 Abstract

The phenylpropanoid pathway is a source of a diverse group of compounds derived from phenylalanine, many of which are involved in lignin biosynthesis and serve as precursors for the production of valuable compounds, such as coumarins, flavonoids, and lignans. Consequently, recent efforts have been invested in mechanistically understanding monolignol biosynthesis, making the quantification of these metabolites vital. The objective of this study was to develop an improved and comprehensive analytical method for (i) extensively profiling, and (ii) accurately quantifying intermediates of the monolignol biosynthetic network, using *Arabidopsis thaliana* as a model system. The method based on liquid chromatography coupled with tandem mass spectrometry was used to quantify phenylpropanoid metabolites in Arabidopsis wild-type lines. A pH of 5.3 and ammonium acetate buffer concentration of 2.5 mM resulted in an optimal analyte response across standards. Vortexing at high temperatures (65^{o}C) enhanced release of phenylpropanoids, specifically the more hydrophobic compounds. Ion suppression was estimated using standard spike recovery studies for accurate quantitation. Compared to existing targeted profiling techniques, our method is capable of quantifying a wider range of intermediates (17 out of 22 in WT Arabidopsis stems) at low *in vivo* concentrations (~50 pmol/g-FW for certain compounds), while requiring minimal sample preparation.

2.2 Introduction

Lignin is an aromatic hetero-polymer synthesized by radical polymerization of hydroxycinnamyl alcohol monomers – also known as monolignols – the end products of the phenylpropanoid pathway (Figure 2.1)[6]. This three-dimensional polymer imparts rigidity and strength to plant cell walls, enabling upright growth, and provides mechanical support and hydrophobicity to plant vasculature, facilitating transport of water and nutrients. While essential to plant viability, lignin impedes degradation of plant cell wall polysaccharides into simple sugars during their fermentation, making biofuel production from lignocellulosic feedstock a cost intensive process[7,17,36]. Therefore, the past two decades have witnessed several genetic engineering efforts targeting the phenylpropanoid pathway, especially in *Arabidopsis thaliana,* to alter lignin amount and composition[26,37]. Despite the progress that has been made, several questions pertaining to control and regulation of carbon flux in this pathway remain unanswered. Recent research efforts have hence been directed towards gaining a mechanistic understanding of lignin synthesis[38–41]. These systems biology driven studies call for accurate quantification of metabolites of the pathway[42,43]. In addition, many phenylpropanoids are industrially relevant products such as tannins, flavonoids, coumarins, and hydroxycinnamic acid conjugates[19], further justifying their quantitation.

MS-based detection associated with various separation techniques (gas and liquid chromatography, capillary electrophoresis etc.) has been widely employed in quantitative metabolic profiling, specifically in analyzing plant metabolomes[44,45]. Reversed phase liquid chromatography (RP-HPLC) coupled to electrospray ionization mass spectrometry is known for achieving high selectivities and sensitivities[44,46,47].

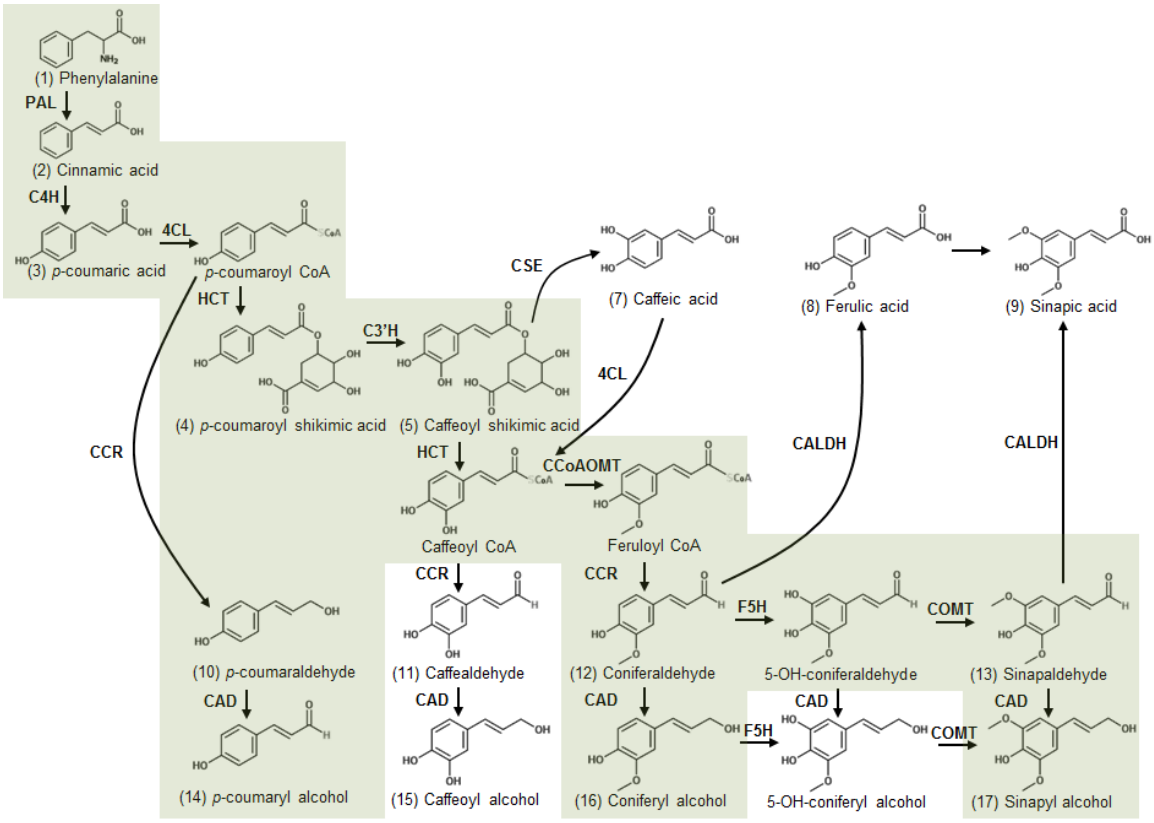


Figure 2.1: Schematic of the phenylpropanoid pathway leading to monolignol synthesis. The highlighted part of the pathway is considered to be most predominant. Intermediates and enzymes currently known in lignin formation are indicated. 4CL, 4-(hydroxy)cinnamoyl CoA ligase; C3′H, *p*-coumarate 3′-hydroxylase; C4H, cinnamate 4-hydroxylase; CAD, cinnamyl alcohol dehydrogenase; CCoAOMT, caffeoyl CoA *O*-methyltransferase; CCR, cinnamoyl CoA reductase; COMT, caffeic acid/5-hydroxyferulic acid *O*-methyltransferase; F5H, ferulate 5-hydroxylase; HCALDH, hydroxycinnamaldehyde dehydrogenase; HCT, hydroxycinnamoyl CoA:shikimate hydroxycinnamoyltransferase; PAL, phenylalanine ammonia-lyase.

Furthermore, instruments equipped with a triple quadrupole (QqQ) allow for fast measurements, improved sensitivities and precise quantitations by multiple reaction monitoring (MRM) [48–50]. Although there is no dearth of analytical techniques in profiling intracellular intermediates, often the data reported is plagued with inaccuracies, which arise from sample handling starting from extract preparations to extract analysis on a detector. Accurate and reliable data can be obtained by addressing and improving various

avenues in LC-MS based analytical methods such as (i) chromatography, (ii) sample preparation and extraction protocols, and (iii) matrix effects. The last decade has seen many studies on profiling phenolics – particularly the ones associated with the phenylpropanoid pathway[51–65]. Although widely employed, these analytical methods have scope for improvement in one or more of the following: (i) simplifying sample preparation, (ii) shortening method run times without compromising compound separation, (iii) accounting for signal suppression because of matrix effects, (iv) canvassing a larger range of intermediates in the pathway.

In an effort to address the aforementioned issues and taking into consideration the sources of error that occur during sample handling and analysis, we developed a rapid, sensitive, reproducible and an accessible analytical method for the accurate quantitation of the metabolites of the monolignol synthesis pathway in Arabidopsis. The effects of chromatographic conditions, such as optimal pH, buffer concentration, column temperature, etc. on analyte responses were investigated. Sample preparation and extraction protocols were tuned to efficiently extract soluble intermediates from *A. thaliana* stem tissue. Ion suppression caused by matrix effects was evaluated using spiking plant extracts with a known concentration of standards. Finally, the analytical method was applied to profile stems from *A. thaliana CCR1* T-DNA insertional lines (*ccr1*) and wild-type (WT) plants.

2.3 Materials and Methods

2.3.1 Chemicals used

L-phenylalanine (>99%), *trans*-cinnamic acid (>99%), *p*-coumaric acid (>98%), caffeic acid (>98%), ferulic acid (>99%), sinapic acid (>99%), shikimic acid (>99%),

coniferyl aldehyde (>98%), sinapaldehyde (98%), coniferyl alcohol (98%), sinapyl alcohol (80%), *p*-fluoro-DL-phenylalanine (>98%), ammonium acetate (>98%) and HPLC grade acetonitrile (ACN) were purchased from Sigma Aldrich (St. Louis, MO). *p*-Coumaraldehyde, *p*-coumaryl alcohol, caffeoyl alcohol and caffealdehyde were synthesized at Discovery Park, Purdue University (West Lafayette, IN). *p*-Coumaroyl shikimate and caffeoyl shikimate esters were acquired from Prof. John Ralph at the University of Wisconsin-Madison (Madison, WI). Glacial acetic acid (>99.7%) was purchased from Mallinckrodt Chemicals (Phillipsburg, NJ) while HPLC-grade methanol was purchased at Macron Fine Chemicals (Center Valley, PA). Water used for making mobile phase solutions was purified using a Barnstead Nanopure Infinity ultrapure water system. All chemicals were used without further processing or purification.

2.3.2 Plant Material

Columbia-0 and the *ccr1* mutant Arabidopsis plants were grown in growth chambers (West Lafayette, IN) at 23^{o}C under 16/8 hour day/night conditions and light intensity of 100 μE m^{-2} s^{-1}. Stems used for analysis were harvested from 5-week old plants. The T-DNA mutant *ccr1* (SALK_123689) was obtained from the Arabidopsis Biological Resource Center. Homozygous *ccr1* mutant was isolated by PCR with primers cc2550 (5′-GTG TCG TAG AGG CTT TGC TTG-3′), cc2551 (5′-TTG TGG AAA TAT TTC CGG TTG-3′), and cc2449 (5′-ATT TTG CCG ATT TCG GAA C-3′).

2.3.3 Standard Solutions

The core monolignol biosynthetic pathway has 22 compounds (Figure 2.1) of which standards for 17 of them were available and considered in the current study. Stock solutions of standards for calibration and determination of limits of detection (LODs) were prepared at a concentration of 0.5 mg/ml in methanol. In the case of shikimic acid, 50/50 (%v/v) methanol-water solution was used owing to its immiscibility in methanol at that concentration. Standard mixtures containing all 17 available compounds were prepared at six different concentrations, approximately ranging from 50 nM to 500 μM, for calibration. *p*-fluoro-DL-phenylalanine was used as an internal standard (IS). All extraction solvents used for the study were prepared with a known concentration of the IS.

2.3.4 Extraction and Concentration of Soluble Metabolites

The basal 0-2 cm fragments of *A. thaliana* stems were harvested and frozen in 2 ml eppendorf tubes using liquid nitrogen. Each biological replicate contained four Arabidopsis stems, which allowed i) to obtain higher yields of the secondary metabolites and ii) to reduce biological variation. To determine the most suitable extraction solvent composition, different concentrations of methanol in water were employed keeping the sample preparation procedure the same.

Stem tissue was pulverized using a pestle in a 2 ml eppendorf tube and 10 μl of solvent was added for every mg of fresh weight (FW) of tissue harvested. The extraction solvents were prepared with the IS at a concentration of 0.001 mg/ml to account for extraction recoveries. The samples were then vortexed for 30, 60 or 120 minutes, using a Midwest Scientific Benchmark Multi-Therm shaker (Valley Park, MO) followed by

centrifugation at 18,000 rpm for 15 minutes. The supernatants from each sample were dried under a vacuum at 30^{o}C using a LABCONCO centrifugal evaporator (Kansas City, MO). The residues were re-dissolved in 60 µl of the extraction solvent and transferred to a standard HPLC vial. Subsequently, 10 µl was injected into the HPLC/MS/MS system for analysis.

2.3.5 Ion Suppression

The extent of ion suppression was quantified using spike recovery method. The dried residues (obtained after extraction from biomass) were reconstituted in 60 µl of the extraction solvent and divided into two parts. One was spiked with 50% (v/v) methanol solution and the other was spiked with a stock solution containing a known concentration of available standards. The study was conducted at three different concentrations of standard compounds in the stock, namely 2, 3 and 5 fold of the concentrations observed in *A. thaliana* stem extracts before accounting for matrix effects. The sample spiked with the extractions solvent (**S**), sample spiked with the stock solution (**SSt**), and the standard stock solution (**Std)** were individually injected into the HPLC/MS/MS system for analysis. The recovery factor (f_i) for each metabolite was computed using equation (Eq. 2.1), where $\boldsymbol{A_{SSt}}$ is the integrated area of a compound in the samples spiked with the stock solution, $\boldsymbol{A_S}$ is the area of a compound in the sample spiked with the extraction solvent, $\boldsymbol{A_{Std}}$ is the area of a compound in the standard mixture, *i* indicates a specific intermediate and 2 is the dilution factor. The analysis was done in triplicate. It should be noted that the recovery factors were estimated *after* taking into account the recovery of the IS added prior to extraction.

$$f_i = 2 \times \frac{A_{SSt_i} - A_{S_i}}{A_{Std_i}} \quad \text{Eq 2.1}$$

2.3.6 Metabolomics using LC-MS/MS.

Chromatography was performed on an HPLC-20AD system from Shimadzu (Columbia, MD) comprising of a quaternary pump, an autosampler, a thermostat controlled column compartment, and a photo diode array detector. Chromatographic separations were performed on a Zorbax Eclipse C8 column (150 mm × 4.6 mm, 5 µm, Agilent Technologies, Santa Clara, CA) at a column temperature of 30ºC and a flow rate of 1ml/min. The injection volume was set to 10 µl. A linear gradient of aqueous solvent A (2.5 mM ammonium acetate in water, adjusted to pH 5.3 using glacial acetic acid) and organic solvent B (98% acetonitrile, 2% water and 0.02% formic acid) was used as follows: 10% B (v/v) for 1 min, 10-20% B over 3 min, 20-20.8% B over 9 minutes, 20.8-50% B over 1 min, 50-70% B over 1 min, hold at 70% B for 3 min, return to 10% B over 1 min, and equilibrate for 4 min at 10% B resulting in a total run time of 23 min per sample. The gradient was unchanged for all aqueous mobile phases considered for the study.

Metabolite profiling was performed using a QTrap 5500 triple quadrupole mass spectrometer from AB Sciex (Redwood City, CA), operating in the negative ion mode. The mass spectrometer is equipped with an ESI-TurboIon-spray interface and all data analysis was conducted using Analyst 1.5.1 software. A low pressure of 1.5 x 10^{-5} torr was maintained in the QTrap 5500 vacuum manifold as indicated by the pressure gauge. The source parameters for the MS were set as follows: curtain gas flow rate, 25 l/h; collision gas, low; ion source voltage, -4.5 kV; desolvation temperature, 700 K; ion source gas 1, 60

l/h; ion source gas 2, 40 l/h. ESI parameters for every standard, such as declustering potential (DP), entrance potential (EP), collision energy (CE), and cell exit potential (CXP) were manually tuned (Appendix Table A1.1). Metabolite recoveries were recorded by subjecting standard mixtures to the extraction protocol (Appendix Table A1.2).

2.3.7 Linearity and Sensitivity

Linearity of standard responses was expressed in terms of the correlation coefficient obtained as a result of a linear fit of the peak areas against the concentrations of the metabolites used for the study.

The measure of sensitivity of the analytical technique was reported as limits of detection (LODs) and limits of quantification (LOQs). These are designated as the concentration of analyte injected that would result in a signal-to-noise (S/N) ratio of 3 and 10 respectively[66]. The LOQs along with the correlation coefficients for all the 17 standards are presented in Table 2.1. Other metabolites, such as the hydroxycinnamic acid derivatives were profiled using ESI parameters of the corresponding hydroxycinnamic acid as standards were not available. The putative retention times and confirmed mass transitions for these compounds have been reported in Table A1.1.

2.3.8 Statistical Analysis

Data were analyzed by one-way ANOVA for independent samples using the online calculator on vassarstats.net/ (Vassar College, Poughkeepsie, NY, USA). A p-value < 0.05 was considered as a significant difference. Tukey's HSD test was employed as a post-hoc test to determine the differences between means. Standard Student's t-test was applied to

analyze differences between individual metabolite concentrations of Arabidopsis *ccr1* and WT stems. P-values of 0.003 have been used after applying the Bonferroni correction to establish a significant difference.

Table 2.1: Retention time (RT), mass transition Q1/Q3 (m/z), and limits of quantification (LOQs) data for the phenylpropanoid pathway intermediates[a]

No.[b]	Metabolite	RT (min)	Q1 $[M-H]^-$	Q3$[M-H]^-$	$R^{2,c}$	LOQ[d] (μM)
1	Phenylalanine	2.53	164.0	147.0	0.99	0.04
2	Cinnamic acid	16.2	147.0	103.0	0.98	26.5
3	*p*-coumaric acid	7.13	163.0	119.1	0.99	0.10
4	*p*-coumaroyl shikimate	7.64	319.2	163.1	0.99	0.05
5	Caffeoyl shikimate	6.02	335.2	179.1	0.99	0.01
6	Shikimic acid	1.54	173.0	93.0	0.99	0.18
7	Caffeic acid	5.16	179.0	135.0	0.98	0.02
8	Ferulic acid	7.87	193.1	178.1	0.99	0.26
9	Sinapic acid	8.49	223.1	208.1	0.99	0.15
10	*p*-coumaraldehyde	11.0	147.0	129.0	0.99	0.01
11	Caffealdehyde	7.41	163.0	145.0	0.97	0.30
12	Coniferaldehyde	12.1	177.1	162.0	0.99	0.12
13	Sinapaldehyde	11.6	207.1	192.1	0.98	0.03
14	*p*-coumaryl alcohol	6.87	149.1	131.0	0.99	2.66
15	Caffeoyl alcohol	5.25	165.1	147.0	0.98	0.17
16	Coniferyl alcohol	7.48	179.1	146.0	0.99	0.03
17	Sinapyl alcohol	7.23	209.1	194.1	0.99	0.68

[a] Analysis was performed using an AbSciex QTrap 5500 mass spectrometer coupled to Shimadzu RP-HPLC system.

[b] Metabolite annotation as represented in Figure 1.

[c] Correlation coefficients from linear fits of standard calibration curves covering a concentrations range of ~50 nM-500 μM.

[d] The LOQs are reported as values at which a signal to noise ratio (S/N) of 10 was obtained.

2.4 Results and Discussion

2.4.1 Improving analyte responses by manipulating chromatography.

Chromatographic conditions, such as mobile phases, buffer pH, buffer concentration, solvent flow rate, and column temperature in addition to achieving

metabolite separation also effect metabolite responses when associated with ESI-MS[67]. A flow rate of 1 ml/min is suggested given the column dimension used for the study and ACN was used as the organic buffer due to its high eluotropic nature and low viscosity[68]. As a result, in this study we focused on the effects of buffer pH and buffer concentration on analyte responses. The solvent gradient, as described in the methods section, was optimized for resolution enough to prevent co-elution of multiple compounds in order to minimize their contribution to ion suppression (Figure 2.2). Using MRM mode does not require strict baseline separation of standards as long as the m/z ratios of the parent (Q1) and fragment (Q3) ions are distinct. As a result, all the optimization strategies considered in the following sections were motivated to obtain higher analyte responses instead of improving resolution. Standard mixtures containing all 17 compounds at a concentration of 0.01 mg/ml – which is within the linear range of calibration – were used for the optimization studies.

Effect of mobile phase pH. Mobile phase pH plays an essential role in the dissociation/deprotonation of analytes and also the ionization of the silanol groups of the column support[69], thereby altering the selectivity and retention on the column[70,71]. The buffers in this study were prepared at a pH of 5, 5.3 and 5.6, which fall within 1 unit of the pK_a of acetate thus ensuring effective buffering capacity[72]. The largest change in retention times was observed for the hydroxycinnamic acids in the pathway when their pK_a values (~4.5) were close to the buffer pH.

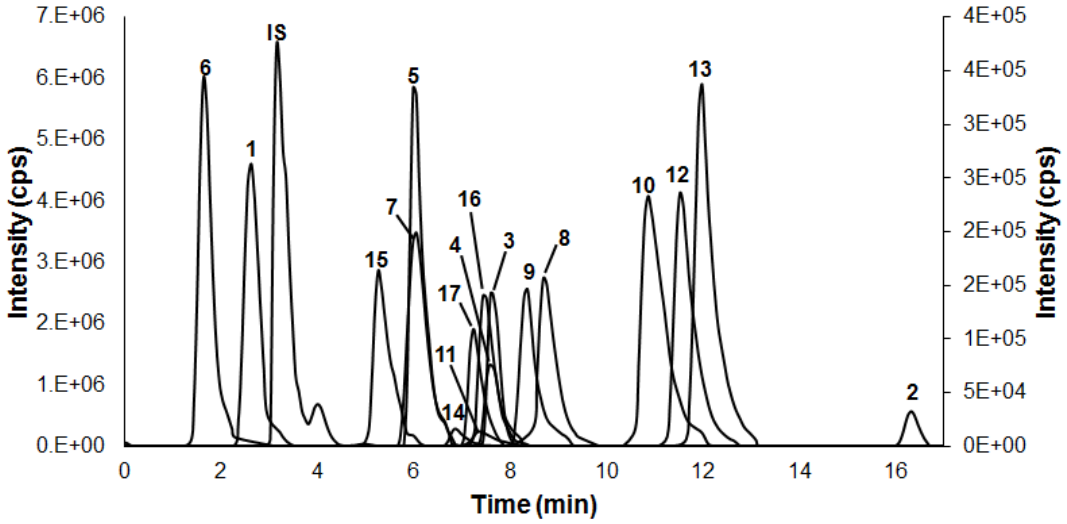


Figure 2.2: Chromatogram obtained by HPLC-MS/MS profiling of a standard mixture of 17 phenylpropanoid metabolites (0.01 mg/ml each). Separation was performed on a Zorbax-Eclipse C8 column (150 mm × 4.6 mm, 5 µm) using 2.5 mM ammonium acetate in water (pH of 5.3) as solvent A and ACN/H_2O/HCOOH (98/2/0.02% – v/v) as solvent B. Two different y-axes were used to accommodate metabolites with very high responses. Compound intensities are reported in counts per second (cps). Metabolites are marked according to Table 2.1

Changing the buffer pH from 5 to 5.6 can cause a four-fold increase in the degree of ionization, as per the Henderson-Hasselbalch relation, thereby altering the retention time. As expected, no significant change in retention time was noted for the extremely polar (phenylalanine and shikimic acid) or hydrophobic compounds (hydroxycinnamyl aldehydes and alcohols). In other words, the pK_a values for the former are too low and the latter are too high compared to the buffer pH such that they either are entirely ionized or neutral. Interestingly, higher responses were observed across all metabolites at the lower pH values of 5 and 5.3 (Figure A1.1). A buffer pH of 5.3 resulted in statistically higher intensities for 11 out of 17 standard compounds relative to pH 5.6 and 6 out of 17 compounds relative to pH 5 (Figure A1.1). Such a relation between mobile phase pH and analyte response in the negative ion mode, termed as wrong-way-around[73], was

previously observed. An acidic mobile phase provides excess protons that are reduced to hydrogen gas during electrospray ionization. The excess negative charges thus accumulate and increase the local pH value eventually aiding in the deprotonation of analytes[74]. All further experiments were performed using a buffer of pH 5.3.

Effect of buffer concentration. The competition between the analyte ion and the buffer anion has also been known to affect the retention factor[69] and sensitivity of the analyte[75]. In order to test the effect of electrolyte concentration on analyte response, standard mixtures were analyzed using buffer solutions consisting of 2.5, 5 and 10 mM ammonium acetate that were adjusted to a pH of 5.3. Higher salt concentration in the mobile phase led to a decrease in retention times and hence lowering the resolution of the analytes. Retention times were not significantly affected but a considerable signal enhancement was observed at buffer concentrations of 2.5 and 5 mM relative to 10 mM across all metabolites (Figure A1.2). A signal enhancement of at least 1.5 fold for all metabolites and up to 3 fold in the case of sinapic acid was obtained at a buffer concentration of 2.5 mM (Figure A1.3). The reduction in signal response at a higher concentration may be due to the competition between the analyte ions and the buffer counter anion during electrospray ionization[75,76]. Buffer concentrations below 2.5 mM were not considered due to a risk of lowering the effective buffering capacity[72] of the mobile phase and increasing the method run time. For all further studies, buffers with salt concentration of 2.5 mM and a pH 5.3 were employed.

Effect of column temperature. A higher column temperature offers several advantages such as (i) faster method runs, (ii) reduced pressure drop, (iii) improved peak shapes[77], and (iv) increased resolution. We performed studies at column temperatures of 30^{o}C and

40^{o}C to observe the effects on analyte resolution and response. No statistical difference was observed in the signal response across all metabolites except sinapaldehyde (**13**, Figure A1.3). The 10^{o}C temperature rise marginally reduced the retention times of most metabolites, while cinnamic acid still eluted at around 16 min (data not shown). Since no significant advantages were incurred using a higher column temperature, all experiments were performed at 30^{0}C.

2.4.2 Improving extraction of soluble phenylpropanoids from *A. thaliana* stem tissue.

Efficient solid-liquid extraction of soluble metabolites is governed by many factors, such as extraction technique used, solid-to-liquid ratio, tissue size, solvent composition, temperature, extraction duration, number of repeated extractions[78–80]. This necessitates optimization of extraction conditions to achieve complete extraction of desired metabolites. As part of our preliminary studies, we tested several extraction techniques such as vortexing, bullet blending and ultrasonication on pulverized Arabidopsis stems at room temperature for a fixed duration and observed no statistical differences (data not shown). Vortexing has been used in all the previously discussed experiments as it is gentle on metabolites and ensures constant suspension (mixing) of plant tissue. A solvent-to-tissue ratio of 10 µl/mg was used for extraction, which is sufficient for standard sample preparation in targeted metabolomics[81]. As a result, we focused mainly on optimizing the extraction solvent composition, temperature, and duration of extraction.

Effect of extraction solvent composition. Composition of an extraction solvent has an inevitable intrinsic bias towards certain metabolite classes given the vast chemical diversity of the plant metabolome[82]. Consequently, it is necessary to use a solvent system that

maximizes the number and amount of metabolites extracted. Previous studies have shown that methanol-water solutions best meet the demands of a chemically heterogeneous system such as the phenylpropanoid pathway[83–85]. We therefore investigated the effects of methanol concentration on metabolite extraction using 50% (v/v) MeOH in water (Control), 75% MeOH in water (M75), or double extraction with pure methanol followed by a wash in 50% (v/v) MeOH in water (MD). Extraction was carried out for 60 minutes at room temperature (25ºC). The one-way ANOVA analysis on 4 replicates in each case resulted in no significant effect of methanol concentration on the extraction of phenylpropanoid metabolites (Table S3). Only 9 metabolites were detected above their limits of quantitation as a result of the extraction (Figure A1.4). Although no statistical variations were observed, 75% (v/v) MeOH in water was chosen as the solvent for all subsequent experiments to ensure deactivation of plant enzymes since fresh tissue is used for extraction.

Effect of extraction temperature. Higher temperatures favor solute dissolution into the extraction solvent and increase solvent accessibility to plant tissue due to a reduced viscosity[79] but there is a trade-off due to certain metabolites being labile to high temperatures. Thus, to test the effect of temperature on metabolite extraction, samples were vortexed at 4, 25 and 65ºC for 60 minutes. One-way ANOVA analysis indicated a significant effect of temperature on metabolite extraction (Table A1.3), with higher temperatures favoring metabolite extraction. Extraction at 65ºC resulted in almost a 10-fold improvement in coniferaldehyde and sinapaldehyde pool sizes, while coniferyl and sinapyl alcohols, sinapic acid and *p*-coumaraldehyde showed 2-4 fold improvements compared to that at room temperature (Figure 2.3, Figure A1.5). It should be noted that the more hydrophobic metabolites of the pathway showed the most improvement at higher

temperatures. The compounds being more hydrophobic may preferably partition into the cell membrane than into methanol-water solution. Higher temperatures may improve analyte solubilities in methanol-water mixture as well as disintegrate cellular membranes enhancing their release into the solvent. In light of these findings, extraction in all further experiments was carried out at 65°C.

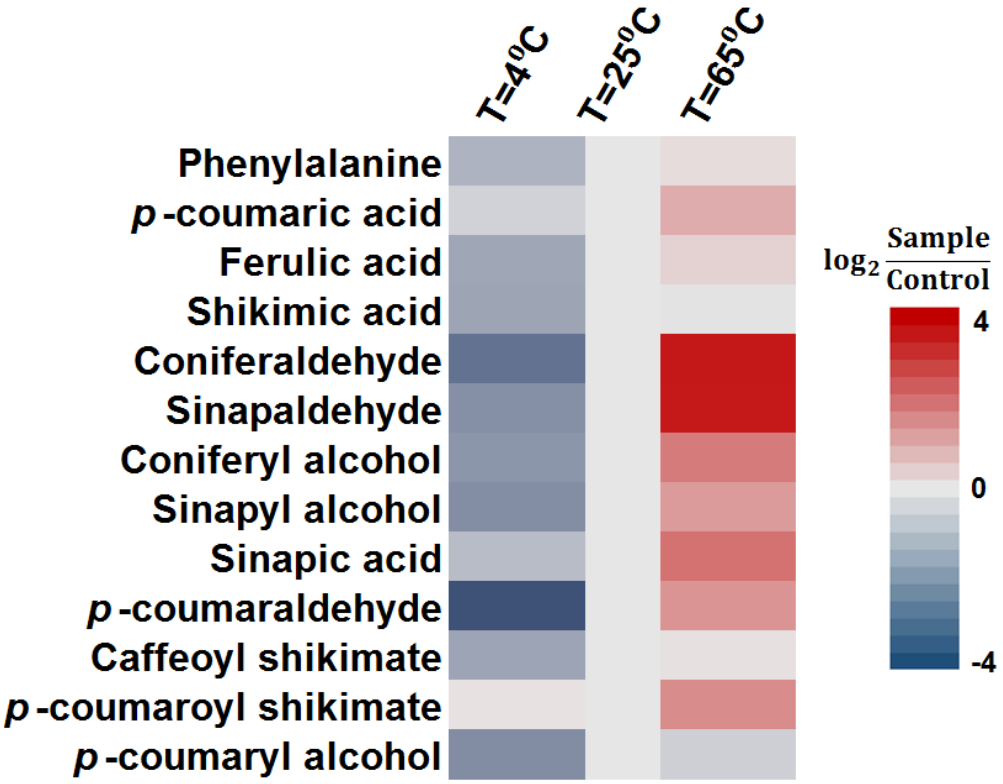


Figure 2.3: Heat map depicting metabolite fold changes as a result of extraction temperature. Data presented as $\log_2$(abundance in sample/abundances at 25 °C). Data are fold changes from (n=4 biological replicates).

Effect of extraction duration. The amount of analyte released may also be a function of extraction time if its extraction is kinetically limited[86]. To investigate this, stem tissue was vortexed at 65°C for 30 min (ED30), 60 min (ED60), and 120 min (ED120). Our results revealed no statistical effect of duration of extraction on increasing metabolite yields (Figure A1.6, Table A1.3). This indicates that the extraction of phenylpropanoids is not

limited by its dissolution kinetics in the solvent and vortexing at a high temperature, as 30 min suffices complete extraction of metabolites. Accordingly, all further sample preparations were conducted by vortexing tissue at 65^{o}C for 30 min. It should be noted that there might be interaction effects considering the number of parameters optimized in this study. Capturing such interactions may lead to an improved response in the form of LODs and LOQs. These are the subjects of future work.

2.4.3 Ion Suppression due to matrix effects.

Tandem mass spectrometry, albeit offering crucial advantages for compound quantitation like high selectivity, sensitivity and throughput, finds its Achilles heel in matrix effects[87]. The alteration in ionization efficiency of an analyte at the electrospray interface due to a co-eluted or co-extracted compound(s) is termed a matrix effect. This phenomenon causes analyte signal suppression leading to incorrect quantitation of compounds of interest. Although the exact mechanism of matrix effects is still debated, it is largely believed to originate because of a competition between analyte and co-eluting matrix components during electrospray ionization. These matrix components may be endogenous species (extracted from biomass) or mobile phase additives[88]. Various approaches have been proposed to minimize ion suppression or account for its effects[89]. One is reduction of injected sample volume or the dilution of the samples, but this hinders the ability to detect certain metabolites. Another approach is reduction of ion suppression by choosing an appropriate sample preparation procedure such as protein precipitation, solid phase extraction, liquid phase extraction etc. Often these methods lead to a decrease or an increase in matrix effects, loss of analytes during extraction and hinder high

throughput analysis of biological samples. Addition of an internal standard[90], either structurally similar to an analyte or labeled with stable isotope, that co-elutes with the analyte can account for losses due to signal suppression. However, this may fail or prove to be expensive when profiling multiple compounds spanning a wide range of chemical properties. Taking all factors into consideration we studied the matrix effects by a standard spike recovery method[91], also known as post extraction addition[92], as described in the methods section. The experiments were conducted with three different spike solution concentrations, namely x2, x3, and x5 fold of the endogenous concentrations of metabolites.

The study was done in 4 replicates and recovery factors (f_i) for all metabolites were determined accordingly. A lower f_i value is indicative of a significant suppression in signal due to matrix effects, while a value close to 1 indicates almost complete recovery of the spiked metabolite. Data from Arabidopsis WT extracts showed no statistical trend of the recovery factors of phenylpropanoid metabolites across different concentrations of the spike solution (Figure 2.4). Shikimic acid suffered the highest signal suppression (f_i = 0.2), while sinapic acid, caffealdehyde, and *p*-coumaryl alcohol resulted in recoveries between 70-80% (Figure 2.4). Reliable recovery factors could not be estimated in case of cinnamic acid due to the spike solution concentrations being close to its LOD, and hence haven't been reported. All the remaining pathway intermediates were almost fully recovered.

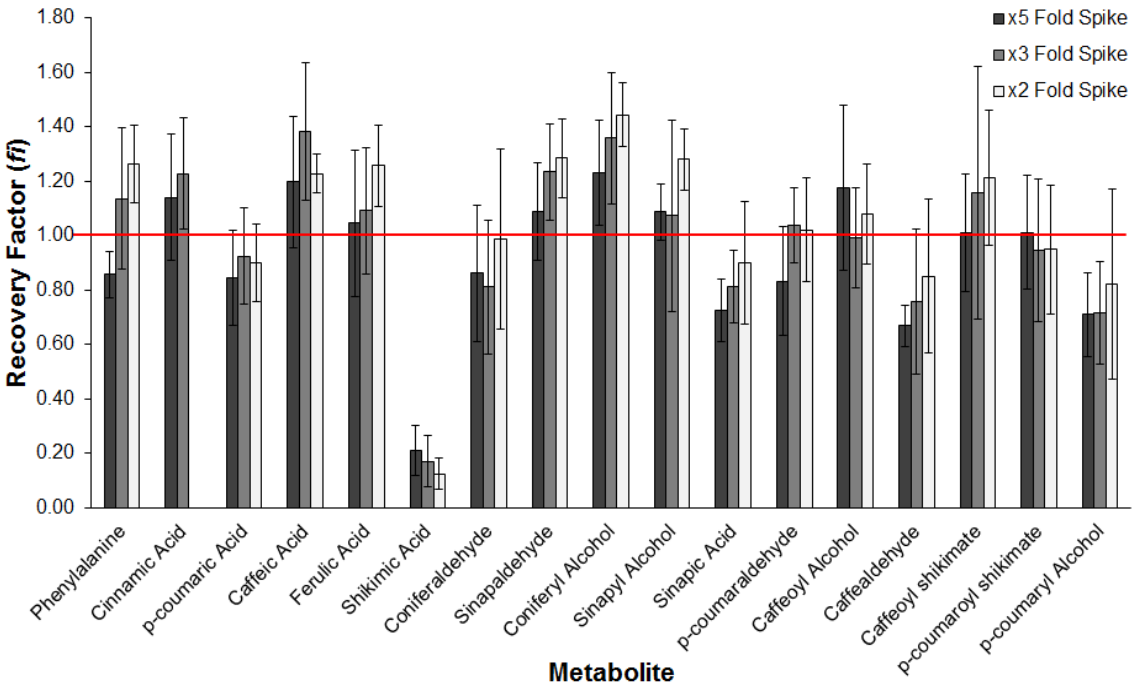


Figure 2.4: Ion-suppression recovery factors obtained by spiking the tissue extract with stock solution containing all standards at 2, 3 and 5 fold of their endogenous concentrations. Data are means ± s.d. (n=4 biological replicates). * = $p < 0.05$ and ** = $p < 0.001$ obtained by Tukey's HSD post ANOVA test.

Shikimic acid, due to its highly polar nature, elutes first from the column at 1.64 minutes (Table 2.1), close to the residence time of un-retained metabolites. This causes it to elute with a host of other endogenous polar metabolites extracted from the plant that may compete for ionization at the electrospray interface. As a result, only 20% of the added shikimic acid was recovered from the spiked sample implying that the true shikimic acid concentrations may be 5 fold higher. Improved chromatographic conditions, individually optimized ESI parameters and use of negative ion mode[93,94] have been conducive in obtaining close to no signal suppression for a majority of the metabolites analyzed.

2.4.4 Metabolite profiling of Arabidopsis WT and *ccr1* lines.

CCR (EC 1.2.1.44) catalyzes the NADPH-dependent conversion of hydroxycinnamoyl-CoA esters to their respective aldehydes, a crucial step in monolignol biosynthesis (Figure 2.1, [19]). Two Arabidopsis enzymes, CCR1 and CCR2, have been kinetically characterized showing high affinity to feruloyl-CoA and lower affinities for sinapoyl- and caffeoyl-CoA[95,96]. It has been previously shown that CCR1 has a greater catalytic efficiency in converting feruloyl-CoA to coniferaldehyde and is primarily involved in lignin synthesis due to its high expression in stem tissue. CCR2, on the other hand, is barely detectable under normal growth conditions and is hypothesized to be involved in pathogen induced lignification[95]. CCR1 knockout plants exhibit a dwarf phenotype, with collapsed xylem vessels, significant reduction in total lignin content and change in composition[97,98], and altered cell wall cohesion leading to improved saccharification efficiency[35]. Given the drastic phenotype invoked as a result of the T-DNA insertion and that CCR1 is required for the production of monolignols, we proposed to profile *ccr1* to visualize changes in the metabolite abundances within the phenylpropanoid pathway. The CCR1 deficient lines were analyzed using the optimized analytical technique and compared with WT Arabidopsis plants.

Our analysis of stem tissue from *ccr1* lines revealed a significant increase in pools of the hydroxycinnamic acids, *p*-coumaric acid (~35 fold), caffeic acid (~12 fold), sinapic acid (~2 fold), with the highest increase observed in ferulic acid (~200 fold, Figure 2.5, Table S4) with respect to the WT stems. Simultaneously, a marked reduction in pools of the hydroxycinnamyl aldehydes and alcohols was detected, with *p*-coumaraldehyde and *p*-coumaryl alcohol approaching their LODs (Figure 2.5, Table A1.4). In addition, we saw

an increase in hydroxycinnamic acid derived esters, such as sinapoyl glucose and feruloyl glucose, in agreement with previous attempts in profiling *ccr1* mutant lines[97,99]. Although ferulic and sinapic acid synthesis occurs via the action of aldehyde dehydrogenases on their respective aldehydes (Figure 2.1), it was interesting to see high levels of these acids when their aldehyde precursors have been depleted. This may be reconciled given that a large increase in the caffeic acid pool can invoke ferulic acid synthesis via COMT by outcompeting (K_m = 24.2 µM[100]) the other substrates. In addition, the feruloyl-CoA esters, accumulated as a result of the knockout, can be hydrolyzed to ferulic acid by the action of putative thioesterases[97]. A part of the ferulic acid so formed can be further hydroxylated by F5H (K_m = 1 mM[101]) to 5-hydroxyferulic acid followed by its methylation via COMT (K_m = 31.6 µM[100]) to sinapic acid (Figure 2.1). Given the concentration of ferulic acid observed (Table A1.4), it is likely that the synthesis of sinapic acid via the suggested route may occur in spite of ferulic acid's low binding affinity to F5H.

Apart from the hydroxycinnamic acid derived esters, other phenylpropanoid derivatives, such as certain kaempferol glucosides were also previously shown to accumulate in *ccr1*[97]. This is reasonable given the large accumulation of *p*-coumaric acid, a precursor to flavonoids. Overall, these results strongly suggest a shift away from lignin synthesis to that of hydroxycinnamic acid-esters and other secondary metabolite derivatives of the phenylpropanoid pathway, in stems of *ccr1*.

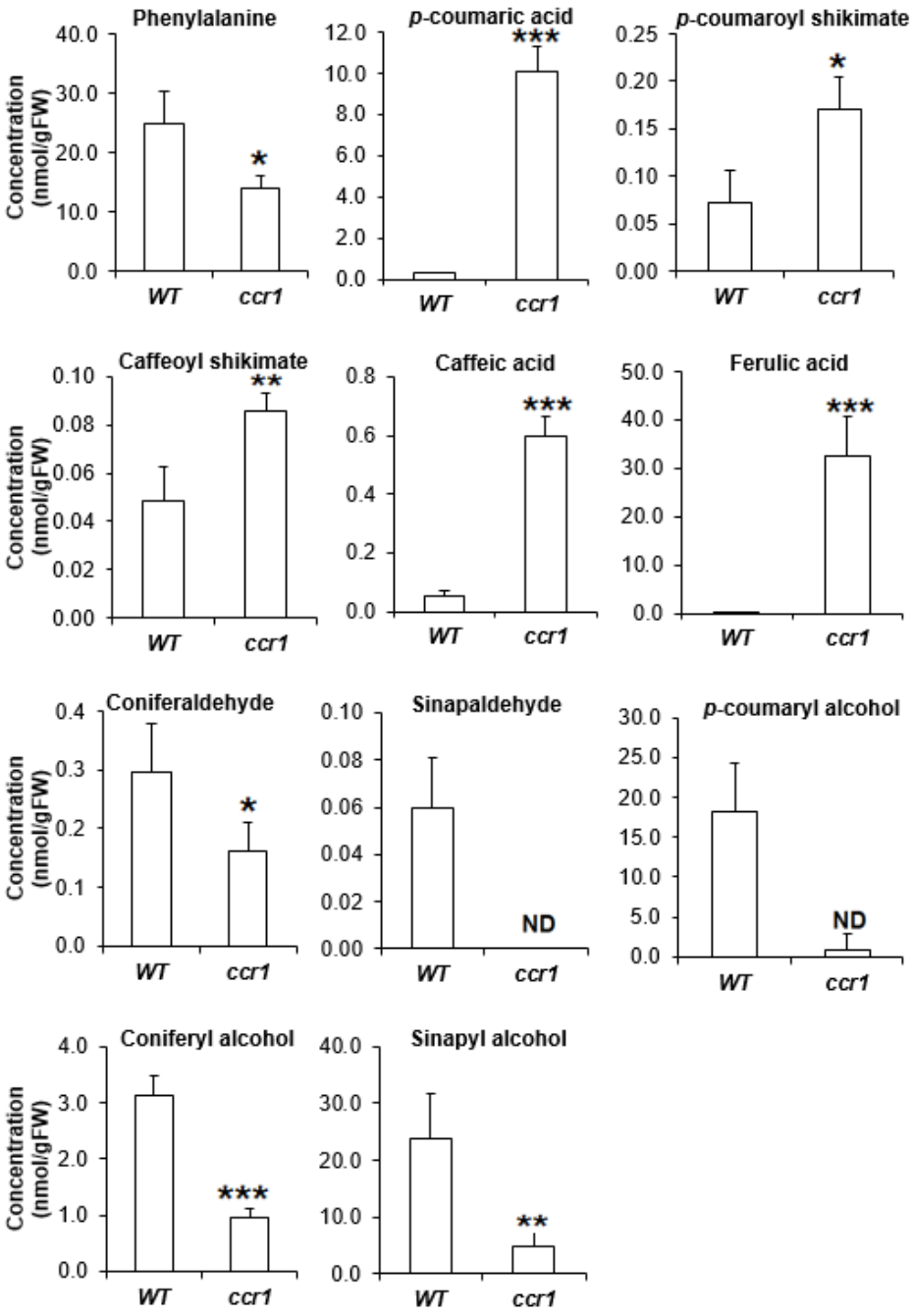


Figure 2.5: Pool sizes of phenylpropanoid pathway intermediates in WT and *ccr1* lines of *A. thaliana* stem tissue. Data of metabolites presented as means ± s.d. (n=4 replicates). Analyte responses normalized to fresh weight of tissue. * indicates $p<0.05$, ** indicates $p<0.001$, *** indicates $p<0.0001$ obtained using the standard Student's t-test. P-value established after the Bonferroni correction is 0.003, indicating that metabolites marked as ** and *** are significantly different. Only metabolites with significant differences between the two lines have been reported in the figure.

2.5 Conclusions

To summarize, we developed a novel, comprehensive LC-MS/MS based analytical method for quantifying intermediates of the phenylpropanoid pathway. In this study, we presented a systematic strategy that optimized every unit operation starting from sample preparation to compound detection by MS. Manipulating chromatographic conditions resulted in a 1.5 to 5-fold increase in analyte responses across standards considered for the study with a buffer pH of 5.3 and buffer concentration of 2.5 mM being the optimal. Extraction studies showed that vortexing at high temperature results in higher yields of analytes. Although no significant effect of the solvent composition or the extraction duration have been observed, this is highly dependent on the metabolites of interest as well as the model system. Quantifying signal suppression due to matrix effects indicate a considerable loss of the shikimate signal due to its co-elution with a host of other polar endogenous metabolites. Applicability of our method was corroborated by quantifying phenylpropanoid intermediates in WT and *ccr1* lines of *A. thaliana*. Our findings were congruent with previous studies profiling WT stems, and to our knowledge, this is the first study presenting absolute concentrations of phenylpropanoid pathway metabolites in *ccr1* lines of Arabidopsis. Of the 17 metabolites in the core metabolite network that have been considered, this study reports accurate *in vivo* concentrations of 15 compounds in Arabidopsis stems, higher in comparison to previous reports of 7[60] and 8[63] metabolites quantified. In addition, this method is able to detect hydroxycinnamic acid derivatives like sinapoyl glucose, sinapoyl malate, feruloyl glucose, and feruloyl malate, and allows for conducting stable isotope labeling experiments demonstrating its potential application in detecting products of enzyme assays, hydrolysis of cell-wall bound phenolics, lignin degradation[58], and systems biology efforts in profiling genetically engineered plants.

3. ANALYTICAL METHOD DEVELOPMENT (II): QUANTIFYING HYDROXYCINNAMYL COENZYME-A THIOESTERS.

3.1 Abstract

In plants, a significant proportion of carbon fixed by photosynthesis is directed toward the phenylpropanoid pathway for lignin synthesis. Hydroxycinnamoyl coenzyme A (CoA) esters are key intermediates and branch points of the phenylpropanoid network. Although CoA thioesters are ubiquitously found in all living systems, they are highly labile and generally accumulate to very low *in vivo* concentrations making their accurate measurement very challenging. In this study, we have developed a novel and facile analytical method based on reversed phase liquid chromatography (RPLC) coupled with tandem mass spectrometry (MS/MS) that has been applied to quantify hydroxycinnamoyl CoA esters in *Arabidopsis thaliana*. The method entails a simple extraction protocol, a short method run time (13 min/sample), and offers almost a 10 to 60-fold improvement in metabolite specific sensitivity compared to the most recent published technique.

3.2 Introduction

3.3 Material and Methods

3.3.1 Chemicals

p-coumaroyl CoA, caffeoyl CoA, and feruloyl CoA were enzymatically synthesized by Prof. Chapple's Lab (Purdue University, West Lafayette-IN). Benzoyl CoA (>90%) was from Sigma Aldrich (St. Louis, MO). Glacial acetic acid (>99.7%) was from Mallinckrodt Chemicals (Phillipsburg, NJ) while HPLC-grade methanol was from Macron

Fine Chemicals (Center Valley, PA). Water for the mobile phases was purified using a Barnstead Nanopure Infinity ultrapure water system. All chemicals were used without further processing or purification.

3.3.2 Plant Material

Arabidopsis Columbia-0 ecotype plants were grown in growth chambers under a 16/8 hour day/night cycle at 23^{o}C and a light intensity of 100 µE m^{-2} s^{-1}. The basal 0.5-2 cm fragments from 5 week old inflorescence stems were used for the analysis.

3.3.3 Standard Solutions

The hydroxycinnamoyl CoA thioesters were purified to a concentration of around 1 mM by chromatography post enzymatic synthesis. Stock solutions for the CoA esters, including Benzoyl CoA, were prepared at a concentration of 500 µM. All stock solutions were stored at -80^{o}C for long term use and -20^{o}C for daily use to prevent degradation. For the purposes of this study, benzoyl CoA was used as the internal standard (IS) because (i) no detectable endogenous pools were found in Arabidopsis stems (data not shown); (ii) has degradation kinetics similar to the CoA thioester intermediates of the phenylpropanoid pathway (data not shown). Standard mixtures containing all 4 metabolites were prepared at six different concentrations ranging from 100 nM to 150 µM.

3.3.4 Stability Studies

Standard mixtures at a concentration of 150 µM from stock solutions stored at -80^{o}C. Three sets of triplicates were prepared with one set at -20^{o}C, one set at 4^{o}C, and the

other left at room temperature (~25^{o}C). Samples were analyzed 24 and 48 hours after incubation at the respective temperatures. The peak areas were recorded to determine analyte stability and to design an extraction protocol amenable to a more sensitive analysis of the CoA esters.

3.3.5 Extraction and Concentration of Soluble Metabolites

The basal 0-2 cm segments of 5-week old *A. thaliana* inflorescence stems were harvested by cutting using liquid nitrogen. The stems were pulverized to fine powder in liquid nitrogen using a mortar-pestle. Each sample contained ~200 mg FW to extract detectable concentrations of the CoA esters. Extraction solvent (75% MeOH in water) containing the IS at a concentration of 0.001 mg/ml was added to the powdered tissue in the ratio of 10 µl to every mg-FW[102]. Samples were then vortexed at 4^{o}C for 30 minutes in a Midwest Scientific Benchmark Multi-Therm shaker (Valley Park, MO). The samples were then spun down in a micro-centrifuge equilibrated at 4^{o}C at 18000 g for 15 minutes. The supernatants were dried under a stream of nitrogen gas and the remaining pellet was re-dissolved in 60 µl of 50% MeOH in water. Samples were then transferred to an HPLC vial for subsequent analysis on the LC-MS.

3.3.6 Metabolomics using LC-MS/MS

Analytes were separated on a Zorbax Eclipse C8 column (150 mm × 4.6 mm, 5 µm, Agilent Technologies, Santa Clara, CA) using an HPLC 20AD system from Shimadzu (Columbia, MD) at a column temperature of 30^{o}C and a flow rate of 1ml/min. The injection volume was 10 µl. A linear gradient of aqueous solvent A (5 mM ammonium acetate in

water, adjusted to pH 6.2 using glacial acetic acid) and an organic solvent B (98% acetonitrile, 2% water and 0.02% formic acid) was used as presented in Table 3.1, resulting in a separation of the here hydroxycinnamoyl CoA ester and sample run time of 13 mins (including equilibration)

Table 3.1: Mobile phase gradient for analyzing hydroxycinnamyl CoA esters.

Time (min)	Solvent A(%)	Solvent B(%)
1	90	10
7	10	90
10	10	90
11	90	10
13	90	10

Metabolite profiling was performed on an AB Sciex QTrap 5500 triple quadrupole mass spectrometer (Redwood City, CA), operating in the negative ion mode. The mass spectrometer is equipped with an ESI-TurboIon-spray interface and all data analysis was conducted using Analyst 1.5.1 software. A low pressure of 1.5×10^{-5} torr was maintained in the QTrap 5500 vacuum manifold as indicated by the pressure gauge. The source parameters for the MS were set as follows: curtain gas flow rate, 20 l/h; collision gas, medium; ion source voltage, -4.5 kV; desolvation temperature, 700 K; ion source gas 1, 60 l/h; ion source gas 2, 60 l/h. ESI parameters for every standard, such as declustering potential (DP), entrance potential (EP), collision energy (CE), and cell exit potential (CXP) were manually tuned to obtain high sensitivities (Table 3.2)

Table 3.2: Retention time (RT), ion transitions Q1/Q3 (m/z), and ESI parameters for the phenylpropanoid pathway intermediates[a]

Metabolite	RT (min)	Q1 $[M-H]^-$	Q3 $[M-H]^-$	LOQ (nM)	DP (volts)	EP (volts)	CE (volts)	CXP (volts)
p-coumaroyl CoA	4.26	912.3	408.1	440	-260	-8	-48	-15
Caffeoyl CoA	4.04	928.3	408.1	1030	-260	-8	-50	-17
Feruloyl CoA	4.36	942.3	408.1	160	-260	-8	-50	-15
Benzoyl CoA	4.48	870.3	408.1		-254	-2	-56	-12
CoA-SH	2.08	766.2	408.1		-200	-12	-43	-21

[a] Analysis was performed using an AbSciex QTrap 5500 mass spectrometer coupled to Shimadzu RP-HPLC system.

3.4 Results and Discussion

3.4.1 Separation and MRM of CoA Esters

Chromatographic conditions, such as mobile phases, buffer pH, buffer concentration, solvent flow rate, and column temperature are crucial to analyte separation, and sensitivity when analyzed using an ESI module [67]. A flow rate of 1 ml/min is suggested given the column dimension used for the study and ACN was used as the organic buffer due to its high eluotropic nature and low viscosity[68]. A buffer pH of 6.2 was chosen compared to the mobile phase (pH 5.3) used for profiling phenylpropanoids as (i) CoA esters are known to be relatively more stable in solutions close to a pH of 7[103,104] and (ii) it is within the range of the buffering capacity of acetic acid[72]. The solvent gradient (Table 3.1) was optimized for obtaining resolution enough to prevent co-elution of CoA esters (Figure 3.1).

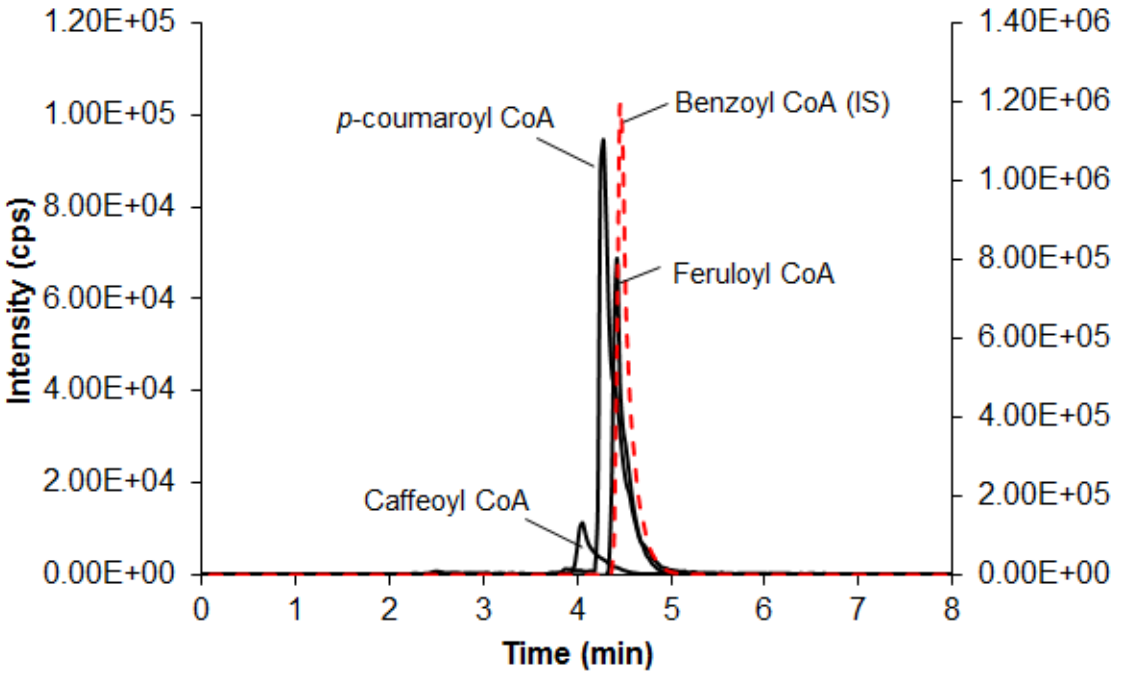


Figure 3.1: Chromatograms of hydroxycinnamoyl CoA thioesters at a concentration of 100 µM. Separation was performed on a Zorbax Eclipse C8 column (150 mm × 4.6 mm, 5 µm) using 5 mM $NH_4CH_3CO_2$ buffer in water (pH 6.2) as solvent A and ACN/H_2O/HCOOH (98/2/0.02 %v/v) as solvent B. Data for feruloyl CoA and benzoyl CoA have been plotted on the secondary axis (right).

The multiple reaction monitoring (MRM) mode offers a unique advantage of monitoring ion transitions (parent, Q1; and fragment/daughter, Q3) making it a highly selective and sensitive technique without requiring a complete baseline separation of the analytes. The parent ion (Q1) for all CoA esters corresponded to the molecular weight after a loss of hydrogen ($[M-H]^-$). However, the fragment ion (Q3→408.1) generated was common to all CoA esters. Loss of metaphosphoric acid and a water molecule from 3'phosphoadenosine disphosphate moiety of Coenzyme A results in the formation of the fragment ion (Figure 3.2, [105]). Having a common fragment ion independent of the acyl moiety allows for 'blind' metabolite profiling of other similar CoA esters in plants (e.g. Sinapoyl CoA, Cinnamoyl CoA etc.). Standard mixtures at a concentration of 100 µM –

well within the linear response of the MS detector– were used for all MS optimization and tuning studies. Metabolite specific ESI parameters and source parameters were manually tuned to obtain higher sensitivities. Our developed chromatography method compounded with tailored tuning of MS parameters resulted in metabolite specific sensitivity gains of over 10 to 60 fold compared to a recent analytical method published on profiling hydroxycinnamoyl CoA esters[106].

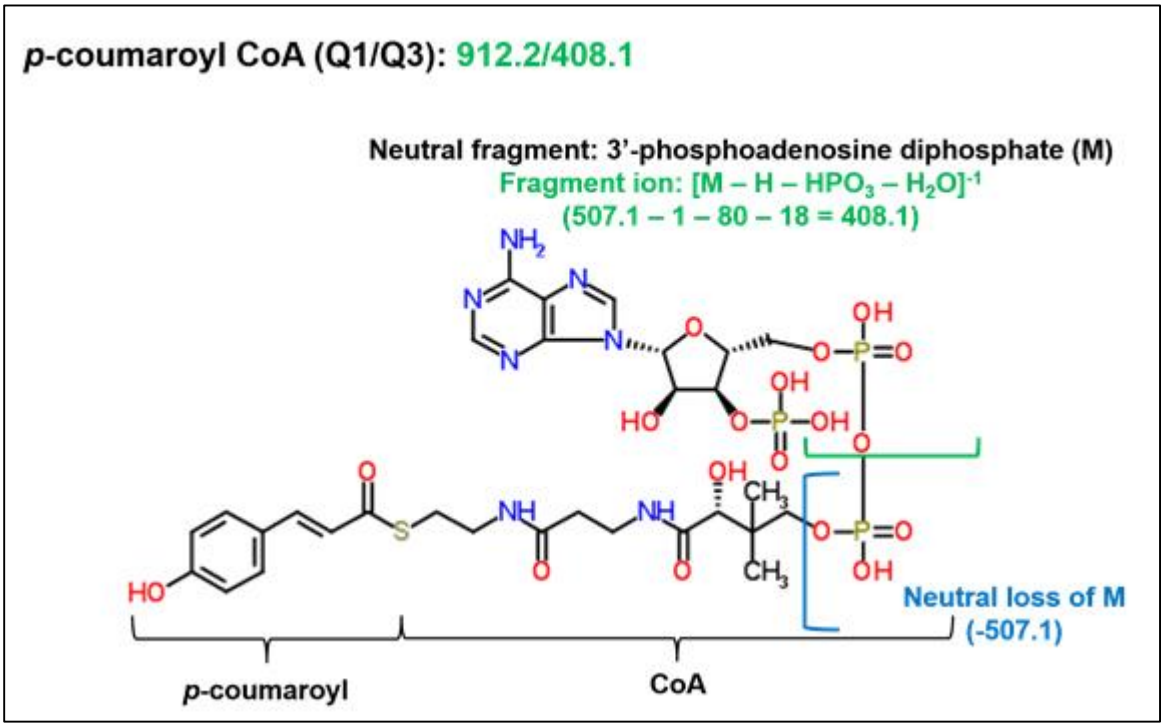


Figure 3.2: Schematic of the ion transition for *p*-coumaroyl CoA (Molecular weight: 913.2 g/mol).

3.4.2 Stability Studies on CoA Esters

Thioesters are known to be very reactive and have a high negative Gibbs free energy of hydrolysis[107]. Certain acyl CoAs are known to degrade by almost 50% of their initial concentration in a few hours at room temperature[103]. To investigate such losses due to

degradation and to design an appropriate sample preparation protocol for extraction studies on actual plant tissue, stability studies were conducted by incubating 100 µM standard mixture of CoA esters at -20 ºC, 4 ºC, and 25ºC (room temperature). Standard mixtures were analyzed 24 and 48 hours after incubation at the respective temperatures (Material and Methods). Almost all CoA esters in the study were stable at -20ºC for the period of two days (Figure 3.3). No significant differences in analyte responses were observed after 24 hours at 4ºC but almost 70-80% of the analyte was degraded by the end of 48 hours. The CoA esters were very labile at room temperatures exhibiting significant losses within 24 hours of their incubation at 25ºC (Figure 3.1). Our findings strongly suggested (i) the use of a representative internal standard that accounts for analyte losses due to degradation, and (ii) to maintain the samples under cold conditions throughout the extraction process during extraction and sample processing right until the time of analysis. Benzoyl CoA was not detectable in WT Arabidopsis stem extracts and is relatively close to the hydroxycinnamoyl CoA esters in terms of its chemical structure. Isotopically labeled analogues of CoA esters would serve as ideal internal standards. To minimize losses due to degradation, all samples were vortexed and spun down at 4ºC and were stored in a -20ºC freezer maintained when not in use. Concentration under a stream of nitrogen was deemed amenable as sample temperatures would be significantly below room temperatures due to evaporative cooling.

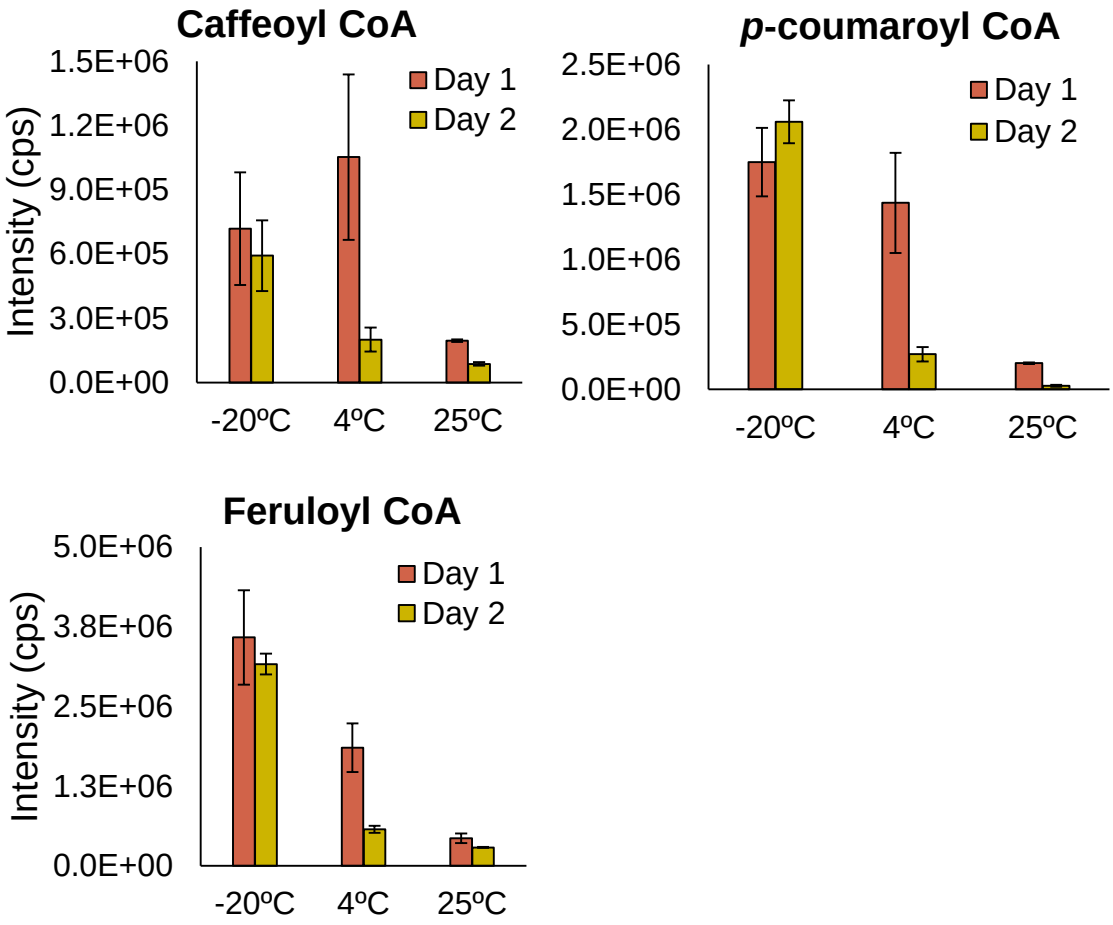


Figure 3.3: Analyte responses from stability studies conducted on CoA ester standard mixtures at a concentration of 100 μM. Data presented as means and standard deviations of intensities from n=3 replicates.

3.4.3 Analyzing Arabidopsis stem extracts

Basal sections of Arabidopsis inflorescence stems were chosen for the study as they are highly lignifying, with increased expression of enzymes of the monolignol biosynthesis pathway[102]. Phenylpropanoid pathway intermediates are also more likely to accumulate in these basal fragments. The entire sample preparation procedure entailing extraction by vortexing (30 mins), centrifugation (15 mins), concentration using a nitrogen evaporator (90 mins), and dissolution of dried extracts (15 mins) was a total of ~3.5 hours, significantly shorter and simpler than some of the previous analytical methods[104,106,108]. HPLC

vials with the final samples were stored at -20^{o}C until their injection on the LC-MS for analysis. All three CoA esters were observed above LOQs in the stem extracts (Figure 3.4). The high sensitivity of the analytical method allowed for quantification of CoA esters at low concentrations of ~0.05 nmol/g-FW. Internal standard recovery on average was found to be 76 ± 14 % of the total amount added during extraction. Accounting for the losses manifested in the precision of the reported metabolic concentrations with observed standard deviations being less than 25% of the mean values (Figure 3.4).

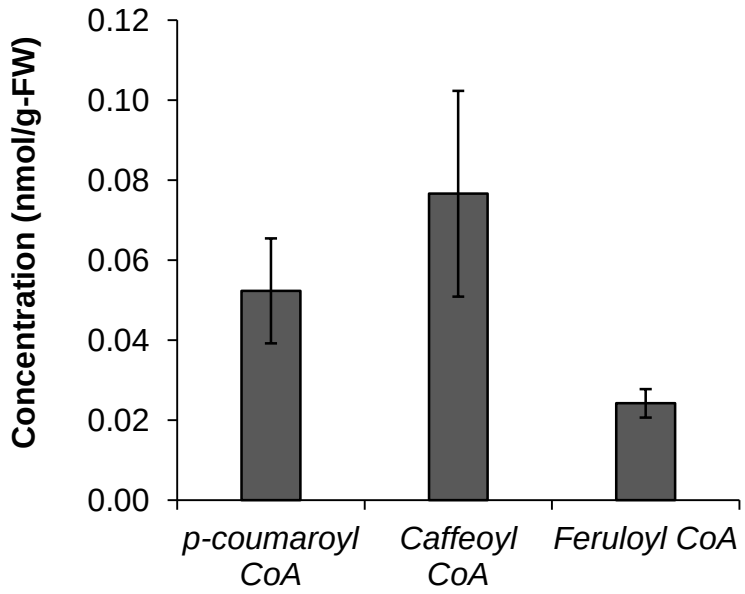


Figure 3.4: Concentrations of hydroxycinnamoyl CoA thioesters in the basal section of 5 week old Arabidopsis WT stems. Data are the means and standard deviations from n=3 replicates.

3.5 Conclusions

In this study, we developed a rapid, adaptable, sensitive, and facile method to quantify hydroxycinnamoyl CoA esters in plant extracts. The chromatographic parameters were conducive in achieving a separation of the four CoA thioesters used in the study. The

analytical technique was found to be over 10 to 60-fold more sensitive than the most recently reported method; for the first time allowing quantification of the hitherto undetectable hydroxycinnamoyl CoA esters in Arabidopsis stems. The ability to measure *p*-coumaroyl CoA, caffeoyl CoA, and feruloyl CoA opens up the study of carbon flux allocation towards lignin at key branch points and potential regulatory sites of the phenylpropanoid pathway. The application of the analytical method to other plant systems would provide new opportunities to investigate phenylpropanoid metabolism.

4. METABOLIC FLUX ANALAYSIS OF THE PHENYLPROPANOID PATHWAY IN ARABIDOPSIS MUTANTS

4.1 Abstract

The phenylpropanoid pathway is highly interconnected with several branch points and is responsible for the synthesis of the three monolignols that constitute the predominant fundamental units of lignin. Consequently, knowledge of the relative contribution of the metabolic pathway fluxes involved in lignin synthesis is essential for rational metabolic engineering of the pathway. In this study, high resolution flux maps of the phenylpropanoid network was computed for in Arabidopsis WT and 4-coumarate ligase knockdown lines (*4cl1*) using ^{13}C-metabolic flux analysis (^{13}C-MFA). Isotopic labeling enrichments and total concentrations of 15 pathway intermediates were measured after exogenously supplying $^{13}C_6$-phenylalanine to stems from both genotypes. Dynamic mass balances were formulated for each metabolite and network fluxes were estimated by fitting a model describing the labeling dynamics to the experimental data. Total acetyl bromide soluble lignin and labeled lignin measurements from DFRC analysis were used as constraints for flux estimation. A reduction in the total incoming flux into the pathway was observed in *4cl1* lines in accordance with the reduced lignin phenotype in the *4cl1* mutant. Although a majority of the incoming flux is still shuttled *via* the traditional shikimate-ester route to lignin synthesis, flux estimations suggested a second route of caffeic acid synthesis from *p*-coumaric acid under fed conditions. The reaction catalyzed by caffeoyl-shikimate esterase (CSE) was also found to significantly contribute to caffeic acid synthesis. A higher flux towards sinapyl alcohol derived lignin (S) was observed in *4cl1* lines with the reaction

catalyzing the conversion of coniferaldehyde to sinapaldehyde is more active than the hydroxycinnamyl alcohol counterpart.

4.2 Introduction

Plants normally channel around 20-30% of photosynthate towards synthesis of the amino acid phenylalanine (Phe) that is further converted to lignin, a hetero-aromatic polymer that imparts structural integrity to the plant vasculature and impedes efficient cellulosic biofuel production [7,109]. Lignin synthesis occurs via the phenylpropanoid pathway where the primary precursor Phe undergoes a series of functional modifications primarily deamination, hydroxylations, and methylations to form *p*-coumaryl, coniferyl, and sinapyl alcohol ([110]; Figure 4.1). These products, also known as monolignols, are transported to the secondary cell walls of plant cells to be polymerized into *p*-hydroxyphenyl (H), guaiacyl (G), and syringyl (S) lignin respectively. Now, with an increasing demand for alternative and renewable sources of energy in light of the rapidly depleting fossil fuels, several metabolic engineering efforts have targeted the phenylpropanoid pathway to investigate the possibility of engineering lignocellulosic feedstock and make them more amenable to pretreatment techniques employed during biofuel production [6,16,26,111–113].

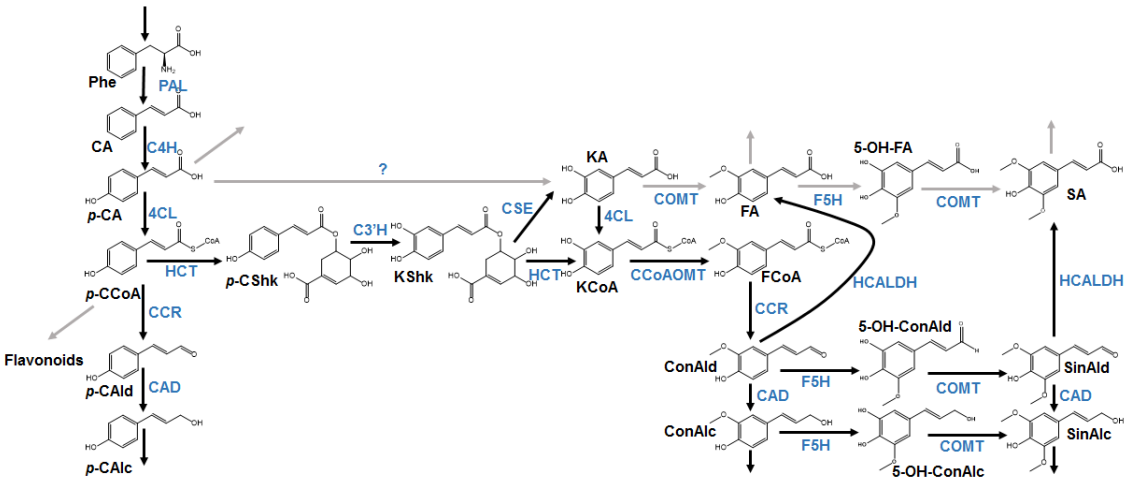


Figure 4.1. Most recent model of the phenylpropanoid pathway leading to lignin biosynthesis. Key reactions are indicated with black arrows. Enzymes are represented in blue and metabolites in black. 4CL, 4-(hydroxy)cinnamoyl CoA ligase; C3′H, *p*-coumarate 3′-hydroxylase; C4H, cinnamate 4-hydroxylase; CAD, cinnamyl alcohol dehydrogenase; CCoAOMT, caffeoyl CoA O-methyltransferase; CCR, cinnamoyl CoA reductase; COMT, caffeic acid/5-hydroxyferulic acid O-methyltransferase; F5H, ferulate 5-hydroxylase; HCALDH, hydroxycinnamaldehyde dehydrogenase; HCT, hydroxycinnamoyl CoA:shikimate hydroxycinnamoyltransferase; PAL, phenylalanine ammonia-lyase.

While some of these experiments successfully reduced lignin resulting in an improved biomass digestibility, several others exhibited a variety of phenotypes leading to dwarfism, plant sterility, slower growth rates, and drastic changes in lignin composition [23,114–117]. Despite significant advances in the technology to manipulate gene expression in higher eukaryotes, genetic engineering in plants often leads to such unanticipated pleiotropic effects due to the sheer complexity of the metabolic networks, sub-cellular compartmentation, and an unpliable regulatory hierarchy. The phenylpropanoid pathway is highly interconnected with many enzymes catalyzing multiple reactions (e.g. 4CL, CCR, CAD, COMT; HCT, Figure 4.1) and several enzymes having multiple isoforms. Moreover, the pathway is characterized by a number of branch points

(e.g. *p*-coumaroyl CoA, caffeoyl-shikimate, coniferaldehyde; Figure 4.1) around which the flux towards the three hydroxycinnamyl alcohols may be manipulated.

In addition to its complexity, in the past three decades the phenylpropanoid pathway has been rewritten several times as seminal studies identified crucial enzymes in the network[118]. The long-standing model of lignin synthesis *via* the hydroxylation of *p*-coumaric acid was reconsidered after concurrent efforts have established that *p*-coumarate 3-hydroxylase (C3´H) – the enzyme responsible for hydroxylation – actively converts 5-*O*-shikimate and 5-*O*-quinate esters of *p*-coumaric acid to the caffeoyl conjugates[115,119]. Furthermore, a recent study has identified the enzyme caffeoyl shikimate esterase (CSE) and established its role in lignin synthesis further amending the metabolic network to now include caffeic acid synthesis via CSE and its conversion to caffeoyl CoA via 4-coumarate ligase (4CL; [25]). What remains elusive is whether the metabolic routes considered in favor of the currently accepted predominant pathway (Figure 4.1) are (i) entirely dormant; (ii) have the ability to be compensatory in the event of downregulation of an enzyme(s) as purported in CCoAOMT downregulation in alfalfa plants where no severe reduction in lignin was observed[120], (iii) become active in transgenic plants leading to a change in the metabolite profiles as seen in CCR downregulated lines where sinapic acid and ferulic acid accumulate even under reduced concentrations of their hydroxycinnamaldehyde precursors[116]. Therefore, despite the vast developments in characterizing the metabolic network, there is a gap in our understanding of the regulation of *in vivo* carbon fluxes and contribution of different routes towards lignin synthesis, which is crucial for proposing rational metabolic engineering strategies.

^{13}C metabolic flux analysis (^{13}C-MFA) is a mathematical technique that quantifies *in vivo* fluxes utilizing isotopic labeling patterns of metabolites[121]. Traditionally applied to microbial systems, MFA is used to assess the effects of environmental and genetic modifications on *in vivo* fluxes, thereby becoming an essential tool in metabolic engineering and systems biology[122–124]. The past decade has witnessed several advances in MFA methodologies and its application to plants, some of them including rice[125], maize[126], soybean[127], *Brassica napus*[128], Arabidopsis[129–134], potato tubers[135,136], and *Petunia hybrida*[137]. Most of these efforts dealt with central carbon metabolism and very little information is available on secondary metabolic networks[135–138].

The main objective of this study is to obtain high resolution flux maps in lignifying stems of Arabidopsis using ^{13}C-MFA and develop key insights into the qualitative and quantitative questions regarding Phe assimilation into lignin. Lines downregulated in 4CL1 were chosen for their interesting phenotype characterized by almost a 30% reduction in total lignin, a higher S to G ratio, but no growth defects[22]. How and why reduced activity of an enzyme catalyzing an early step towards lignin synthesis results in altered lignin composition was an endearing question to further investigate. Isotope labeling experiments were conducted using ring labeled Phe ($^{13}C_6$-Phe) and its incorporation into 15 phenylpropanoid pathway intermediates and lignin was quantified at multiple time points. Flux maps for WT and *4cl1* lines were compared to gain insight into flux splits at major branch points and major flux redistributions as a result of 4CL1 knockdown.

4.3 Materials and Methods

4.3.1 Plant Material

Columbia-0 and *4cl1* mutant Arabidopsis plants were grown in a growth chamber at 23^{o}C under a light intensity of 150 µE/m^{2}-s and a 16/8 hour day/night cycle. The TDNA mutant *4cl1* (WiscDsLox473B01) was obtained from the Arabidopsis Biological Resource Center and confirmed by genotyping PCR (Li et al., 2015).

4.3.2 Isotopic Labeling Study

Primary inflorescence stems from 5-week-old plants were excised under water with a double edged blade and inserted into microcentrifuge tubes containing a solution of 1 mM ring labeled $^{13}C_6$-Phe in Murashige and Skoog (MS) medium. Basal 0.5-2 cm of the stems were harvested at 0 (unfed), 20, 40, 90, and 180 min post feeding, rinsed with water and quenched using liquid nitrogen. The study was done in triplicate using a total of 18 stems for each replicate in order to allocate sufficient tissue for soluble metabolite analysis, total lignin analysis, and lignin composition analysis using derivatization followed by reductive cleavage (DFRC) procedure

4.3.3 Analysis of Soluble Metabolites using LC-MS/MS

Sample preparation and extraction. Frozen stem tissue was ground to a fine powder using a mortar and pestle to which 10 µl of the extraction solvent was added for every mg fresh weight of the harvested tissue. The extraction solvent used for the study was 75% methanol in water with the internal standards benzoyl CoA and *p*-F-(DL)-Phenylalanine at a concentration of 1 µg/ml and 0.1 µg/ml respectively. Benzoyl CoA was

selected as an internal standard for the hydroxycinnamyl CoA esters because, (i) the degradation kinetics are similar to the CoA esters (data not shown), and (ii) there were no detectable endogenous pools in Arabidopsis stem extracts (data not shown). The extraction of the phenylpropanoid metabolites was conducted sequentially in two steps. First, the samples were subjected to a cold extraction by vortexing in a Multi-therm incubated vortexer (Valley Park, MO) at 4^{o}C for 30 min for analysis of the labile hydroxycinnamyl CoA esters. The samples were centrifuged at 15000 g for 15 min and the supernatants (S1) were dried under a stream of nitrogen gas. The second extraction was conducted by adding the same volume of extraction solvent to the remaining pellets followed by vortexing at a temperature of 65^{o}C. The supernatants (S2) were concentrated under a stream of nitrogen gas. Concentrations of hydroxycinnamoyl CoA esters were reported by analyzing S1 samples using LC-MS/MS, while the concentrations of the remaining intermediates of the phenylpropanoid pathway were reported as a combination of S1 and S2 samples.

Analytical methods for metabolite profiling. Analytes were separated using a Shimadzu HPLC 20AD system on a Zorbax Eclipse C8 column (150 mm 4.6 mm, 5 μm, Agilent Technologies, Santa Clara, CA) at a column temperature of 30^{o}C and a flow rate of 1 ml/min. Metabolite detection was achieved using an AbSciex QTrap 5500 triple quadrupole system equipped with an electrospray ionization (ESI) probe in the negative ion mode. Peak areas corresponding to $[M-H]^-$ and $[M+6-H]^-$ ions were integrated to quantify unlabeled and labeled metabolites respectively. Chromatographic conditions and mass spectrometric parameters for phenylpropanoids (other than the hydroxycinnamoyl CoA esters) were set as described previously[102]. Mobile phases and chromatography gradient from Jaini et al., (2017) were altered and optimized for analysis of

hydroxycinnamoyl CoA esters (Table A2.1). The retention times, mass transitions, and optimized ESI parameters for CoA esters have been reported in the appendix (Table A2.2).

4.3.4 Total Lignin Content and Composition Analysis.

Cell wall residue for lignin analysis was isolated as described previously[139]. Mature stems of Col-0 wild type, and *4cl1* plants were harvested and ground to a fine powder in liquid nitrogen. The pulverized tissue was first washed with 0.1 mM sodium phosphate buffer (pH 7.2) at 50 ^{o}C and then extracted with 70% ethanol five times at 65^{o}C. The samples were washed once with acetone and then dried under room temperature. Total lignin content was measured using the acetyl bromide-soluble lignin method described previously[140,141]. For DFRC analysis, 8-15 mg of the dried CWR was dissolved overnight in 2.5 ml of solvent containing acetic acid/acetyl bromide (80/20 %v/v) with 0.2 mg of 4,4'-ethylidenebisphenol as the internal standard (IS). The dissolved samples were dried and redissolved in 2 ml of dioxane/acetic acid/water (50/40/10, %v/v/v). This mixture was then reacted with Zinc dust and the products were acetylated using a pyridine/acetic anhydride mixture (40/60, %v/v). The acetylated lignin derivatives were quantified using gas chromatography FID and mass spectrometry after accounting for the response factors from the internal standard as previously described[142–144].

4.3.5 Mathematical Modeling

Metabolic network and model setup. The phenylpropanoid pathway from Figure 1 was transformed to a network consisting of 15 metabolites and 26 fluxes (Figure 4.2). Cinnamic acid was lumped with *p*-coumarate and considered to be at a steady state as it

was below the limits of detection of the analytical method even after feeding for 3 hours. Reactions involving 5-OH-ferulic acid, 5-OH-conferaldehyde, and 5-OH-coniferyl alcohol were lumped as a single reaction due to lack of standards.

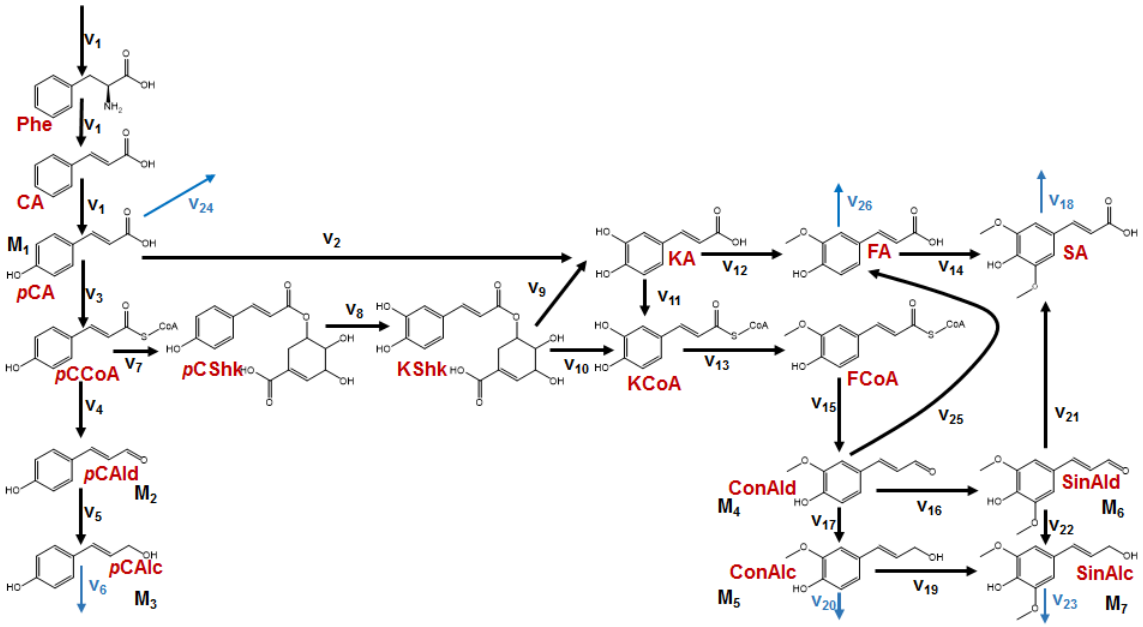


Figure 4.2 Metabolic network for MFA of Arabidopsis. The network consists of 26 fluxes ($v_{1\text{-}26}$). Fluxes to lignin (v_6, v_{20}, v_{23}) and hydroxycinnamic acid derivatives (v_{18}, v_{24}, v_{26}) constitute the exit fluxes of the pathway and are represented in blue. Metabolites for which inactive pools have been invoked are represented as $M_{1\text{-}7}$.

Feeding at a concentration of 1 mM $^{13}C_6$-Phe resulted in a linear accumulation of all phenylpropanoid pathway metabolites. This allowed for expressing the total mass balances on metabolites using a linear equation (Equation 1) with the slope as the difference of incoming and outgoing fluxes ($\Delta\boldsymbol{v}$) and the intercept (**M(0)**) representing the initial metabolite concentrations. Best fit values and confidence intervals for slopes and intercepts for every metabolite for both WT and *4cl1* lines were obtained using linear regression (Table A2.3).

$$M = (v_{in} - v_{out}) + M(0) = \Delta v_i * t + M(0)$$ Equation 1

Component mass balances on labeled metabolites were expressed using ordinary differential equations according to Equation 2, where $\boldsymbol{M_L}$ and $\boldsymbol{M_{UL}}$ indicate labeled and unlabeled concentrations of a metabolite respectively, $\boldsymbol{v_{in}}$ is the input flux from the precursor, $\boldsymbol{v_{out}}$ is the output flux, and $\boldsymbol{f_i}$ indicates the fractional label in a metabolite (Equation 3).

$$\frac{dM_{L_i}}{dt} = v_{in_i} * f_{in_i} - v_{out_i} * f_i$$ Equation 2

$$f_i = \frac{M_L}{M_{UL}+M_L}$$ Equation 3

Framework of Flux Estimation. In summary, MFA entails identifying a set of fluxes that best captures the dynamics of the ^{13}C labeling data. The fluxes were evaluated using non-linear weighted least squares regression where the objective function formulated as the difference between experimentally measured and simulated labeled metabolite concentrations was minimized (Figure 4.3). Inverse of standard deviations obtained from experimental measurements were used as weights for the regression routine. The model consisted a total of 33 parameters for WT and 34 parameters for *4cl1* lines, 26 of which corresponded to the fluxes in the pathways. One parameter was allocated for estimating the actual labeling concentration of Phe (C_{Phe}). C_{Phe} accounts for the dilution in label enrichment in Phe due to the plastidial pool. The remaining parameters represent concentration of inactive pools of *p*-coumaric acid, coniferaldehyde, sinapaldehyde, *p*-coumaryl alcohol, coniferyl alcohol, and sinapyl alcohol. In case of *4cl1*, an inactive pool for *p*-coumaraldehyde was also included due to a lower measured label incorporation than *p*-coumaryl alcohol.

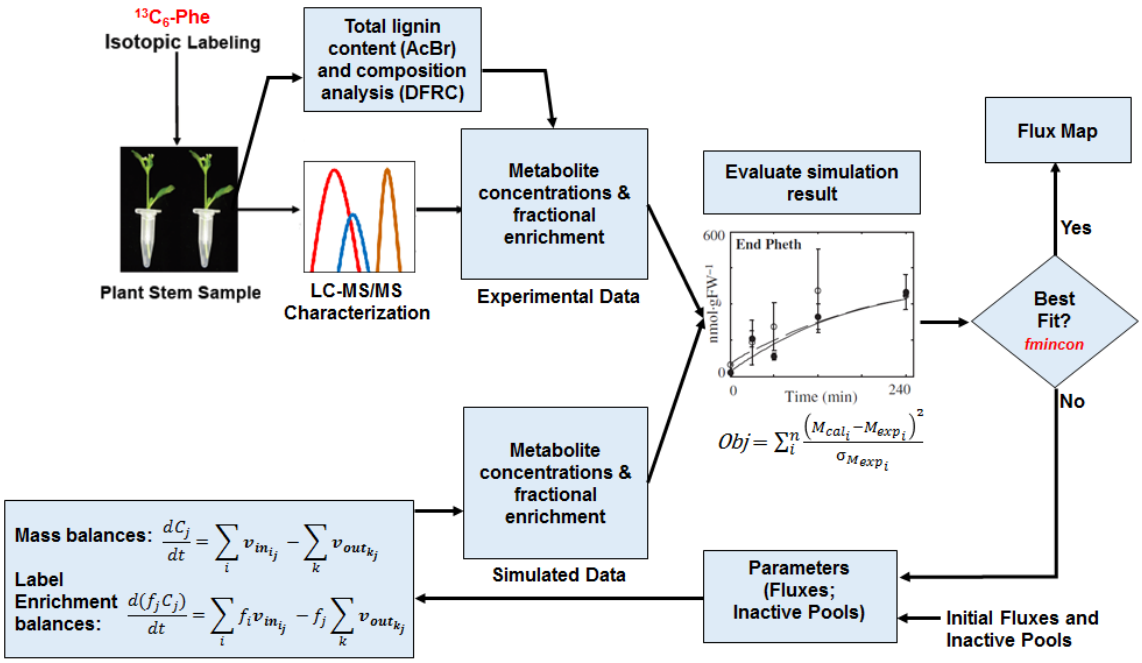


Figure 4.3: Overall framework of the modeling strategy.

The FMINCON function from the Optimization toolbox of MATLAB R2017a was used for all simulations in the study. The rates of change of metabolite concentrations were estimated using linear regression and were used as equality constraints for the optimization routine. Fluxes towards H (v_6), G (v_{20}) and S (v_{23}) lignin were constrained using labeled lignin accumulation data from DFRC analysis (Figures A2.1 & A2.2). Fluxes toward hydroxycinnamic acid derivatives (v_{18}, v_{24}, v_{25}) were equated to the rate of difference in their concentrations before and after hydrolysis (Figure A2.3) for *4cl1* lines. In case of WT, the summation of fluxes towards the hydroxycinnamic acid derivatives was constrained to be less than 10% of the incoming flux based on measurements of sinapoyl derivatives from LC-MS and spectrophotometry analysis.

Estimation of Confidence Intervals on Fluxes. Intercepts and slopes for total metabolite concentrations were randomly sampled using the means and standard deviations estimated from linear regression as previously discussed (Table A2.3). The *normrnd* function was used to carry out the sampling 100 times resulting in a set of 100 metabolite accumulation rates and initial concentrations. The optimization routine was repeated for every sample. The final set of fluxes were reported as the mean and standard deviations from 100 samples.

Hierarchical Clustering Analysis. Clustering analysis was conducted using dynamic isotopic label enrichment data obtained for all metabolites on JMP®, Version 9.2, (SAS Institute Inc., Cary, NC). The Ward's minimum variance method was used to perform the clustering analysis.

4.4 Results and Discussion

4.4.1 Targeted Metabolomics Data across different genotypes.

Stem tissue from both WT and *4cl1* plants were profiled for phenylpropanoid metabolites using LC-MS/MS (Materials and Methods; [102]). A total of 15 metabolites were profiled, of which caffeic acid and caffeoyl CoA were below the quantitation limits in WT and *4cl1* stems respectively. Our analysis revealed a significant increase in the concentrations of hydroxycinnamic acids (Figure 4.4), *p*-coumaric acid (~120 fold), caffeic acid (~250 fold), ferulic acid (~12 fold), and sinapic acid (~2 fold) in *4cl1* lines relative to WT. 4CL1 isoform is known to preferably catalyze the conversion of *p*-coumaric acid and caffeic acid to their respective CoA thioesters. Therefore, high levels of accumulation of the major substrates is expected in *4cl1* plants that exhibit a 70% reduction in enzyme activity[22].

Simultaneously, a significant reduction in the concentrations of *p*-coumaryl- and coniferyl-aldehydes and alcohols was observed (Figure 4.4).

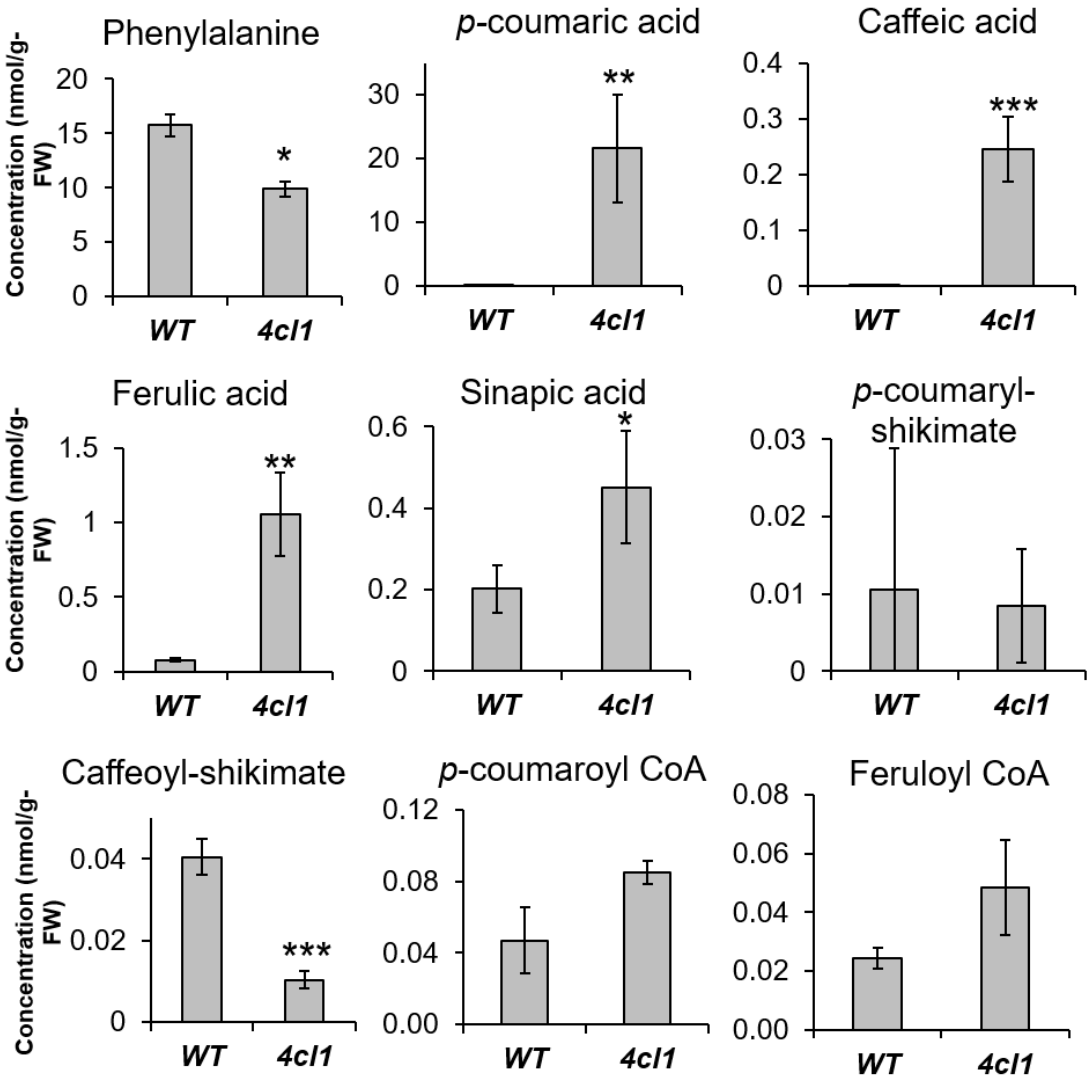
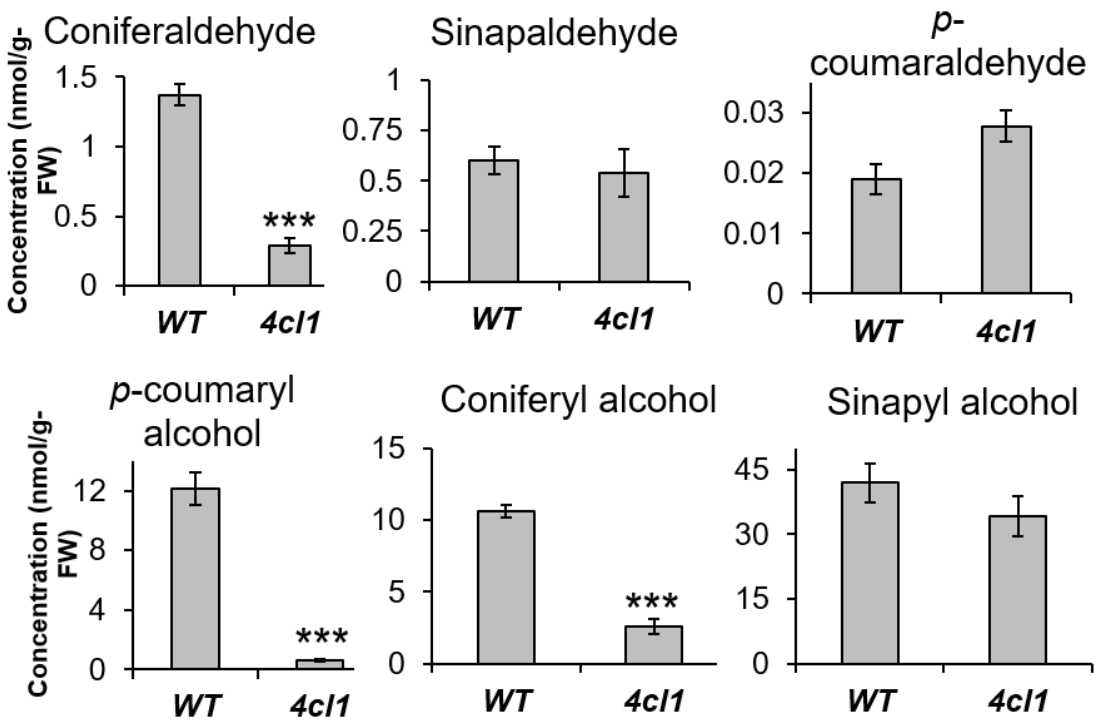


Figure 4.4. Metabolite concentrations of phenylpropanoid intermediates in basal 0-2 cm stem sections of non-fed WT and *4cl1* lines of Arabidopsis. Data presented as mean ± S.D. from n=3 replicates. *p < 0.05, **p < 0.01, and ***p < 0.001 were obtained using standard Student's *t*-test. Data for sinapoyl glucose and sinapoyl malate were normalized to WT measurements for lack of standards.

3.4. Continued



Although ferulate (K_M = 200 µM; [145]) and sinapate (K_M = n.d.; [145])have weaker affinities to 4CL1, an increase in their concentrations may be attributed to the conversion of caffeic acid via COMT and F5H enzymes. This alternative route was previously proposed in light of ferulic acid accumulation in *4cl1* plants with significantly reduced concentrations of its predominant precursor, coniferaldehyde[146]. A marked reduction in the concentrations of sinapoyl derivatives was also observed indicating a reduced flux towards their synthesis in *4cl1 lines* (Figure 4.4). Despite the fact that a reduction in monolignols are in agreement with the reduced lignin levels observed in *4cl1* lines, it is interesting to see a drastic reorganization of the metabolic profile of the phenylpropanoid pathway in plants that are phenotypically similar to WT lines. Nevertheless, it is well-known that metabolite concentrations solely are not informative

about the fluxes through a metabolic network. In order to obtain further insight into the effects of the genetic modification, metabolic flux analysis using $^{13}C_6$-Phe as an isotopic tracer was performed.

4.4.2 Dynamic Labeling Experiments

Isotopic labeling data was obtained for 15 metabolites using the analytical methods described previously. The M+6 isotopologue was successfully detected and quantified for all metabolites profiled over the course of the feeding study, except for Caffeoyl CoA in *4cl1* lines. While upstream metabolites reached an isotopic steady state at the end of the feeding study, an increasing label enrichment was observed for most downstream metabolites, specifically the hydroxycinnamyl aldehydes and alcohols (Tables A2.4 & A2.5). Significant observations from the labeling experiments have been discussed case by case as follows.

Sinapaldehyde and sinapyl alcohol pools in *4cl1* lines have higher ^{13}C enrichments than WT. A label enrichment of ~20% and ~15% was observed for sinapaldehyde and sinapyl alcohol, respectively, in WT lines at the end of the feeding study, while the enrichments were in *4cl1* lines were ~35% and ~19%, respectively. A higher label incorporation in sinapyl alcohol without a change in the endogenous pools indicates a higher synthesis flux in accordance with the higher S lignin phenotype reported for *4cl1* lines[22]. Simultaneously, a decrease in sinapic acid label enrichment from ~59% in WT lines to ~50% in *4cl1* lines was observed without a significant change in the endogenous concentrations. This is indicative of reduced incoming flux from sinapaldehyde, the predominant precursor of sinapic acid, which is in agreement with the reduced

concentrations of sinapic acid derived esters, such as sinapoyl glucose and sinapoyl malate (Figure 4.4; [22]).

Higher fractional label incorporation in products compared to precursors suggests presence of inactive or compartmented pools. The general rule in isotopic labeling studies is that the label enrichment in a product metabolite is always less than or equal to the label enrichment of the precursor. This is a direct eventuality of mass balances on the labeled fraction for each metabolite. Any divergence from this rule essentially violates the law of conservation of mass. However, there are two scenarios where such a discrepancy is still valid, (i) when there exist multiple routes leading to the formation of the product, and (ii) when there exists an inactive pool that is localized either in another cellular compartment, or another cell entirely. For example, the former scenario can explain the higher labeled fraction of sinapyl alcohol in *4cl1* lines when compared to coniferyl alcohol. Sinapaldehyde with a higher labeled fraction than both monolignols is also a precursor to sinapyl alcohol. The latter scenario can be invoked to explain the lower label enrichment of Phe compared to *p*-coumaric acid in WT lines. There could be a significant plastidial pool or a vacuolar pool (Lynch et al., 2017, in press) of Phe that dilutes the labeled fraction of the metabolite when extracted. In case of coniferaldehyde, the inactive pool is suggested to be localized in cellular membranes[102]. Hydrophobic compounds such as hydroxycinnamyl aldehydes and alcohols, having high octanol-water partition coefficients, favorably partition into cellular membranes rendering them inaccessible to the cytosolic enzymes for further conversion[147]. Extraction at high temperatures using organic solvents would release the metabolite partitioned into the membrane thereby reducing the final percentage of labeled intermediate[102]. Thus, inactive pools were invoked for a

number of metabolites for both *4cl1* and WT lines when formulating the model for estimating fluxes in order to sustain the law of conservation of mass (Table A2.6).

A CoA independent route to caffeic acid may be active under fed conditions. Hierarchical clustering of all metabolites measured was performed using dynamic fractional enrichment data for both genotypes. Hydroxycinnamic acids clustered together in both WT and *4cl1* plants indicating that they follow the same labeling dynamics (Figure 4.5 (a)&(b)). The similarity of labeling dynamics is a function of the proximity of the metabolites to each other. In other words, a product should tend to cluster with its immediate precursor. This is evident from the two primary clusters obtained in both genotypes, one that is characterized mainly by upstream metabolites and the second by the more downstream hydroxycinnamoyl aldehydes and alcohols. Interestingly, caffeoyl shikimate – precursor to caffeic acid – clusters with the downstream metabolites. Moreover, the percentage of ^{13}C label incorporation in caffeic acid, ferulic acid, and sinapic acid was found to be consistently higher than their precursors caffeoyl-shikimate, coniferaldehyde, and sinapaldehyde respectively in both WT and *4cl1* lines (Tables A2.4 & A2.5). The most plausible explanation for this would be an alternative route of synthesis for the hydroxycinnamic acid (Scenario 1 from previous section). It is possible, under fed conditions an accumulation in *p*-coumaric acid allows this alternate CoA independent route to caffeic acid. The caffeic acid so formed is in turn converted to ferulic and sinapic acid by the action of COMT and F5H enzymes. Ferulic acid synthesis at higher concentrations of caffeic acid is plausible given the turnover number (k_{cat}/K_M) of COMT for caffeic acid is of the same order of magnitude as the other 5-OH-feruloyl substrates[148].

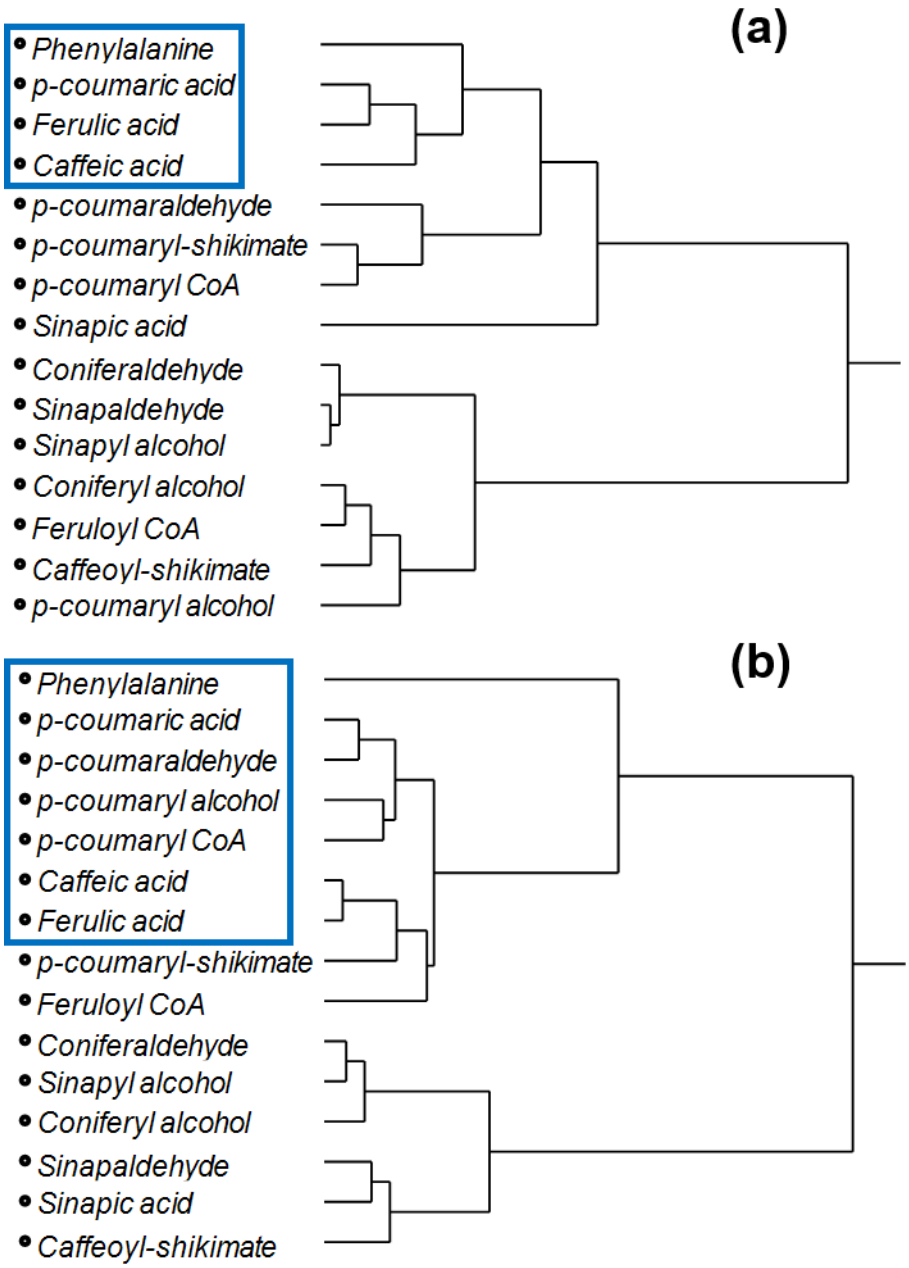


Figure 4.5: Dendrograms obtained from hierarchical clustering of dynamic label incorporation in phenylpropanoid metabolites in (a) WT, and (b) *4cl1* lines. Data from all five time points were included for the analysis. Blue box encloses all hydroxycinnamic acids and the precursor, Phe.

What is interesting and seemingly unlikely is the *in vivo* hydroxylation of *p*-coumaric acid to caffeic acid. *In vitro* enzyme assays expressing C3´H in yeast microsomes

suggested that k_{cat} for *p*-coumaric acid (1.2 nmol/mg-protein/hr; [149]) is around 6000-fold lower compared to that of *p*-coumaryl-shikimate (7344 nmol/mg-protein/hr; [119]). Compounded with a higher K_M of *p*-coumaric acid, the relative catalytic efficiency (k_{cat}/K_M) would be significantly higher than 6000. Despite this fact, caffeic acid formation is still observed in fed stems with *p*-coumaric acid accumulating to ~150 fold in WT stems and ~8000 fold in *4cl1* lines. In other words, *p*-coumaric acid is able to compete with its shikimate ester counterpart at lower concentrations than expected from *in vitro* assays. This brings forth an interesting question as to whether C3´H is the primary enzyme hydroxylating *p*-coumaric acid, or there exists a P450 analog that may have lower expression but is more active towards the hydroxycinnamic acid. Another possibility is the existence of a membrane bound protein complex of P450s that directly converts cinnamic acid to caffeic acid, as was found in case of Poplar[150]. For all simulations conducted in this study, the hydroxycinnamic acid route (fluxes v_2, v_{14}, v_{15}; Figure 4.2) was included in the model in light of the above findings.

Monolignol labeling patterns alone are insufficient and inaccurate to estimate total flux to lignin Total AcBr lignin measurements indicated a ~17% reduction in total lignin in *4cl1* plants (Table A2.7), but no significant changes in lignin deposition during the course of the feeding study for both the genotypes (Figures A2.1 & A2.2). This is contrary to expectations as labeled monolignol units were deposited in lignin as quantified by DFRC analysis (Figures A2.1 & A2.2). The standard errors of measurement of total lignin are of the order of the total flux towards lignin, which makes a confident estimation of lignin deposition rate difficult. Linear interpolation of the DFRC data from fed WT and 4cl1 lines suggests a total lignin (H, G, and S combined) deposition rate to be ~10.3 nmol/g-

FW-min and ~9.9 nmol/g-FW-min respectively. It was interesting to see a similar label deposition rate in *4cl1* lines although it exhibits a reduced lignin phenotype[22]. To investigate this and to obtain constraints for MFA, total flux towards lignin was estimated by fitting parametric curves on label enrichment data of monolignols for both WT and *4cl1* plants (Table 4.1). The total flux towards lignin was estimated to be ~64 nmol/g-FW-min in WT and ~60 nmol/g-FW-min in *4cl1* lines under fed conditions. If this were true, a concomitant increase in total lignin deposition rate should have been observed from the measurements of the AcBr method of quantifying lignin. Despite the fact that the total flux calculated is 2 to 3-fold higher than the variance of the AcBr technique, our measurements indicate no significant accumulation of lignin (Figures A2.1& A2.2).

Table 4.1: Estimates and bounds of fluxes towards lignin.

	Fluxes WT (nmol/gFW-min)			Fluxes *4cl1* (nmol/gFW-min)		
	IE[a]	LB[b]	UB[c]	IE[a]	LB[b]	UB[c]
H	4.5 ± 0.8	1.6 ± 0.2	2.6 ± 0.7	4.5 ± 0.4	2.3 ± 0.4	4.3 ± 0.5
G	45 ± 4.6	7.7 ± 0.9	19.3 ± 1.9	34 ± 5.7	5.4 ± 0.5	12.6 ± 1.6
S	14 ± 2.1	0.9 ± 0.1	3.1 ± 0.3	19.6 ± 1.8	2.3 ± 0.3	5.4 ± 0.4
Total	63.5 ± 7.5	10.2 ± 1.3	25 ± 2.9	58.2 ± 7.9	10 ± 1.2	22.3 ± 2.5

[a]IE: initial flux estimates using measured label enrichments of monolignols
[b]LB: lower bounds of fluxes equaling the rate of accumulate of labeled lignin
[c]UB: upper bounds of fluxes estimated using measured label enrichments of nearest precursor

The possible overestimation of the total flux towards lignin is a result of isotopic dilution due to monolignol pools in the apoplastic space at the time of feeding. In other words, the actual enrichment in the hydroxycinnamyl alcohols is higher than what is measured, which

would in turn reduce the estimates of the total flux towards lignin to a more physiological value. However, the fraction of the monolignol pools localized in the secondary cell walls is unknown, limiting a fixed estimation of the total flux towards lignin. Nevertheless, the labeling data can be used to set bounds on the total flux by eliciting two extremes cases of label enrichment in the hydroxycinnamyl alcohols. The lower bound for the total flux would equal to that of the labeled lignin deposition rate – a scenario where all three monolignol pools are completely turned over (label enrichment fraction = 1). The upper bound of the flux was estimated after equating the label enrichment of the monolignol pools to the nearest precursor for which an inactive pool has not been invoked (Table 4.1). In case of *p*-coumaryl alcohol, this nearest precursor is *p*-coumaryl CoA as there is still a possibility of *p*-coumaraldehyde partitioning into cell membranes. Similarly, feruloyl CoA serves as a common precursor for both coniferyl and sinapyl alcohol.

4.5 Metabolic Flux Analysis

A non-stationary MFA technique was employed to calculate fluxes as the metabolites accumulated during the course of the feeding study. The main assumption for the model was that a step change in the precursor concentration (Phe) results in a step change in the flux values throughout the metabolic network[135,136,151]. This allows us to express mass balances on the metabolites – accumulating at a constant rate – as a difference of constant fluxes (Material and Methods). There have been alternative techniques where the time series data is divided into different metabolic regimes and estimating fluxes for each

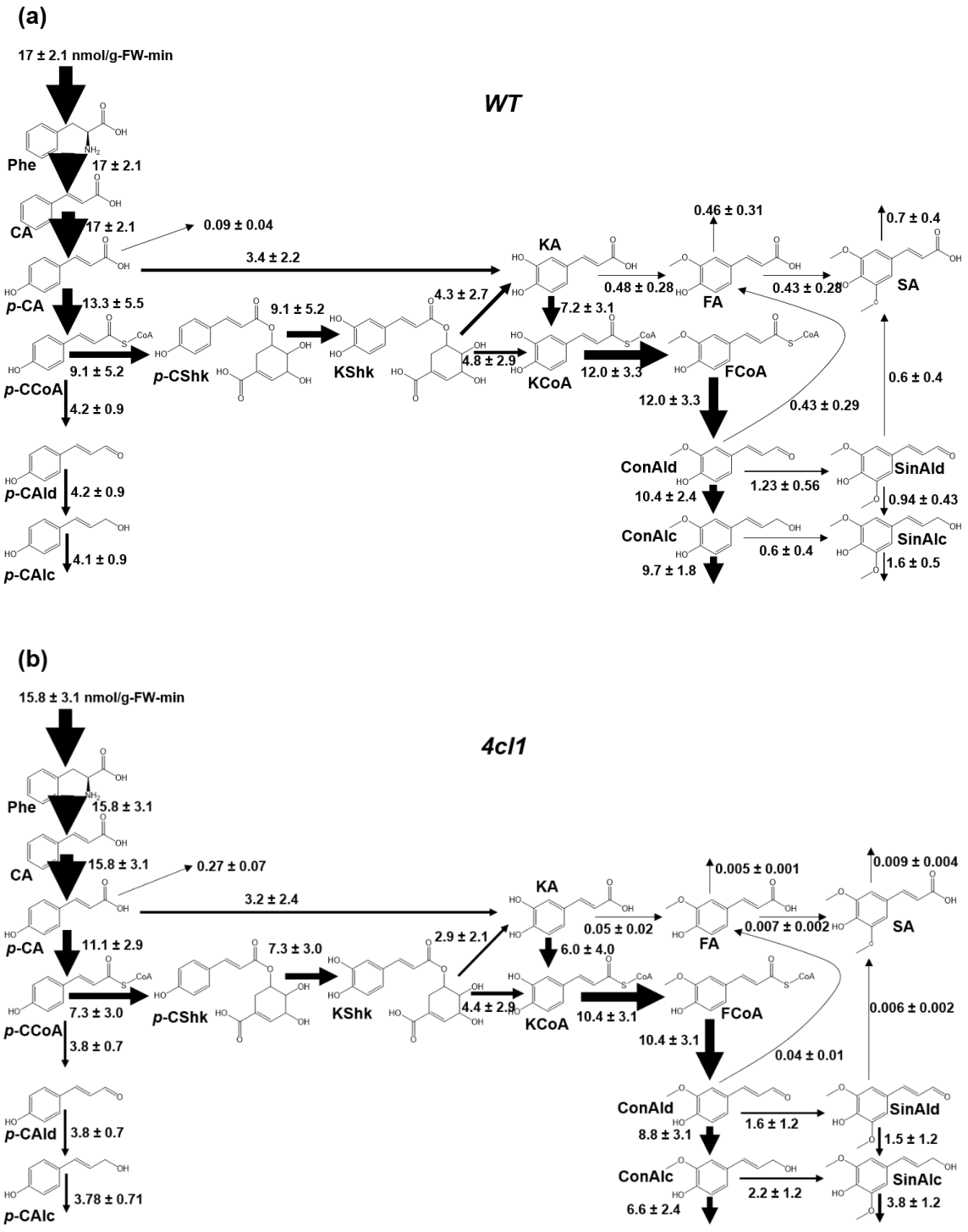


Figure 4.6: Flux maps obtained for WT (a) and 4cl1 (b) lines under fed conditions. Fluxes were represented as mean ± S.D. from (*n=100*) samples obtained by bootstrapping. The thickness of the arrows represents the relative value of the fluxes normalized to the incoming flux (v_1).

regime to capture the temporal evolution in the fluxes[152,153]. More advanced techniques express fluxes as piecewise functions, both linear and non-linear, to describe the temporal profile of fluxes[154,155]. These techniques are very powerful when the flux evolution in a system are gradual, feeding study is conducted over long durations, or wide range of dynamics in metabolite concentrations are observed. Given the high concentration of the label precursor (1 mM) in our study accompanied with moderately sized metabolic network, it was surmised that the constant fluxes assumption would be sufficient to capture the isotopic labeling patterns observed.

The base model for both WT and *4cl1* lines consisted of 26 reactions (fluxes) with inactive pools for *p*-coumaric acid and coniferaldehyde. The final flux maps obtained after feeding 1 mM $^{13}C_6$-Phe in WT and *4cl1* lines are presented in Figure 4.6. Standard deviations were obtained using the bootstrap sampling technique as described in the Materials and Methods section. The set of fluxes represented in Figure 4.6 best fit the experimentally measured labeled metabolite concentrations (Figures A2.4 & A2.5).

Table 4.2 Estimates of inactive pools from the model.

	WT			*4cl1*		
	Inactive Pool (nmol/g-FW-min)		% of total initial pool	Inactive Pool (nmol/g-FW-min)		% of total initial pool
	Mean	S.D		Mean	S.D	
p-coumaric acid	0.15	0.05	84.4	11.5	5.1	53.4
p-coumaraldehyde				0.01	0.00	55.8
p-coumaryl alcohol	5.85	2.64	53.6	0.01	0.01	65.1
Coniferaldehyde	0.67	0.12	90.1	0.11	0.02	86.8
Coniferyl alcohol	7.57	1.33	88.2	0.89	0.36	65.6
Sinapaldehyde	0.28	0.05	89.4	0.13	0.04	76.0
Sinapyl alcohol	30.7	1.85	89.2	18.5	6.37	82.8

4.5.1 Relative fluxes through the reactions catalyzed by 4CL1 are comparable in both genotypes.

Despite a ~10% reduction in the total flux estimated in *4cl1* lines, the flux through the reactions (v_3, v_{11}) catalyzed by 4CL are still significant. Almost 75% of the total input flux passes through *p*-coumaric in both WT and *4cl1* lines, which is reasonable as its conversion to *p*-coumaryl CoA thioester is the first committed step to lignin biosynthesis. Interestingly, flux through caffeic acid was found to be a significant contributor to caffeoyl CoA synthesis constituting almost 45% of the input flux in WT and 37% of the input flux in *4cl1* lines. Such high fluxes in the *4cl1* lines are supported by over a 200-fold accumulation of both *p*-coumaric and caffeic acid (Figure 4.4). Increased concentrations of the substrates may compensate for the reduced activity of the 4CL.

4.5.2 An alternative hydroxycinnamic acid route to Caffeoyl CoA synthesis is active under fed conditions in both WT and *4cl1* lines.

Flux estimations of the hydroxycinnamoyl-shikimate ester route (v_7, v_8) conform with the accepted route (Figure 4.1) with almost 50% of the input flux being shuttled to G and S lignin synthesis in both genotypes. Our findings go hand in hand with the efforts by Bonawitz et al., (2014) where C3´H knockdown lines in the study exhibited dwarfism with lignin composed almost completely of *p*-coumaryl alcohol derived units – establishing the predominant role of the enzyme in lignin biosynthesis. Nevertheless, flux through the parallel route at the hydroxycinnamic acid level was estimated to be ~20% and ~22% of the total flux in WT and *4cl1* lines respectively, indicating the previously reconsidered route to caffeic acid synthesis may still be active[157]. Although this conclusion comes with a caveat that the estimated fluxes are valid when fed with 1 mM Phe and an argument can be made that the alternative route is active only at high concentrations of the substrate that may not be physiological. Indeed, *4cl1* mutant serves as the perfect counter to the aforementioned argument where *p*-coumaric acid accumulates to over ~200 fold in a line that shows no growth phenotype and grows to almost WT height presenting a physiologically valid scenario for alternate hydroxylation of *p*-coumarate to caffeic acid. In other words, although the CoA independent route is not the primary route to lignin synthesis in WT plants, it should be considered while designing genetic engineering experiments and/or analyzing the effects in engineered plants.

4.5.3 Significant flux towards caffeic acid synthesis via CSE.

The phenylpropanoid pathway was recently updated after Vanholme et al., (2013) identified and established the important role of CSE in hydrolyzing caffeoyl-shikimate to

caffeic acid. The authors further went on to show that CSE knockout mutants had drastic phenotypes with almost a 35% reduction in lignin composed of almost evenly distributed H, G and S subunits. Fluxes estimated from our analysis indicate almost a 1:1 split of the reaction fluxes catalyzed by CSE and HCT in WT, and a 1:1.5 split in case of *4cl1* lines. No significant changes in fluxes via CSE and HCT under reduced concentration of caffeoyl shikimate in *4cl1* plants may suggest a change in enzyme levels. Taken together, caffeic acid synthesis from *p*-coumaric acid and caffeoyl-shikimate constitutes almost 40% of the total incoming flux into the phenylpropanoid pathway under fed conditions reiterating the fact that hydroxycinnamic acids other than *p*-coumaric acid are essential intermediates in lignin biosynthesis.

4.5.4 Higher flux towards S lignin in *4cl1* lines supported by estimated fluxes.

Label deposition into S lignin was relatively higher in *4cl1* lines as observed from the DFRC data. This was supported by the estimates of flux to S lignin being ~10% of the total flux in WT and ~25% of the total flux in *4cl1* lines. In addition, flux to the hydroxycinnamaldehyde route was higher in WT consistent with the higher turnover number of F5H with respect to coniferaldehyde(5 pkat/mg-protein-µM; [101]) than coniferyl alcohol(2 pkat/mg-protein-µM; [101]). In case of *4cl1* lines, flux through the hydroxycinnamyl alcohol route was statistically the same as the flux through the hydroxycinnamaldehyde route. This may be due to a significant fraction of the coniferaldehyde pool (~87%) being inactive (Table 4.2, Figure 4.4) while only 65% of the coniferyl alcohol pool was estimated to be compartmented allowing it to compete for F5H. What is intriguing and counterintuitive is that the absolute fluxes in both these branches

are significantly higher in the *4cl1* lines when both coniferaldehyde and coniferyl alcohol pools were reduced by ~5fold (Figure 4.4). This may be indicative of a positive change in enzyme levels in the *4cl1* mutant. Indeed, this hypothesis is in accordance with microarray data of *4cl1* lines of Arabidopsis showing an increased transcript abundance of almost all genes (PAL to CCR) involved in monolignol synthesis[146].

4.6 Conclusions

Using ^{13}C-labeling and flux analysis, high resolution flux maps of the phenylpropanoid pathway were obtained in WT and *4cl1* lines of *Arabidopsis thaliana*. Dynamic labeling experiments using $^{13}C_6$-Phe as the substrate revealed the presence of inactive pools for *p*-coumaric acid in *4cl1* lines, and coniferaldehyde in both WT and *4cl1* lines resulting in isotopic dilution in these metabolites compared to their products. Flux analysis in combination with the label enrichment data indicated the alternative route of hydroxylation of *p*-coumaric acid to caffeic acid is active in both genotypes under fed conditions. However, C3´H is unlikely to catalyze this conversion as *p*-coumaric acid fails to accumulate to a concentration high enough to compete with the predominant substrate *p*-coumaroyl-shikimate, suggesting the presence of an alternative enzyme or a P450 analog responsible for the hydroxylation that remains elusive. Our flux estimates also revealed a significant contribution of CSE to lignin synthesis with almost an even flux split at the caffeoyl-shikimate branch point. Higher flux towards S lignin was observed in *4cl1* lines in accordance with the higher S lignin phenotype compared to WT plants. Flux resolution in the phenylpropanoid pathway could be further improved by profiling 5-OH-ferulic acid, 5-OH-coniferaldehyde, and 5-OH-coniferyl alcohol. The modeling and experimental strategy presented in this study can be used to investigate flux maps in other transgenic

lines of Arabidopsis and other species that exhibit unique phenotypes; to gain insight into phenylpropanoid metabolism and design rational metabolic engineering experiments targeting lignin biosynthesis.

5. TARGETED METABOLOMICS OF THE PHENYLPROPANOID PATHWAY IN ARABIDOPSIS GENOTYPES

5.1 Abstract

The phenylpropanoid pathway is a source of a diverse group of compounds derived from Phe, many of which are involved in lignin biosynthesis and serve as precursors for the production of valuable compounds such as coumarins, flavonoids, and lignans. Consequently, the metabolic network has been a target of many genetic engineering efforts that resulted in a wide range of phenotypes. Metabolite profiling has been widely applied to plants for diagnostic and phenotypic analyses, and with recent advances in analytical techniques it has great potential for directly elucidating plant metabolic processes. In this study, we quantify metabolites of the phenylpropanoid pathway in 5 week old stems of various Arabidopsis genotypes using an analytical technique based on liquid chromatography-tandem mass spectrometry. Multiple reaction monitoring (MRM) was employed to monitor precursor and product ions of metabolites enabling high sensitivity and specificity. A total of 15 intermediates of the phenylpropanoid were quantified across all genotypes. Comparative analyses were performed between genotypes showing significant reorganization in the metabolic profile of the phenylpropanoid pathway.

5.2 Introduction

The phenylpropanoid pathway is a repository of essential plant metabolites derived from the carbon skeleton of the amino acid phenylalanine (Phe) [110]. Compounds of this pathway are referred to as “secondary metabolites” due to their role in plant defense, structural support, and survival[20]. Although the phenylpropanoid pathway is a source of

precursors to many valuable groups of chemicals like flavonoids, anthocyanins, coumarins and other phenylpropanoid derivatives, it is indispensable to plants for its role in the synthesis of hydroxycinnamyl alcohols, commonly known as monolignols. Monolignols constitute the fundamental units of the hetero-aromatic polymer lignin, that imparts structural support and vascular integrity to plants.[158] While crucial to plant sustenance, lignin is one of the major impediments to efficient biofuel production as it renders useful lignocellulosic feedstock recalcitrant to biochemical and mechanical pretreatment techniques, hence lowering the polysaccharide yields[26].

Consequently, several enzymes of the phenylpropanoid pathway are targets of genetic engineering attempts to reduce lignin content and alter its composition for improved forage digestibility[7,8,26,159,160]. Although some of these efforts resulted in lignin phenotypes favorable for saccharification, some resulted in severe to moderate phenotypes that had unchanged if not poorer saccharification efficiencies. Many of these studies have raised questions regarding carbon flux control and regulation in the lignin biosynthesis pathway. In this vein, systems biology approaches are increasingly being employed in mechanistically understanding metabolic networks and visualizing how individual pathways are interconnected[41,137,161–165]. Although integrative ‘omics’ approaches are favored, metabolomics is said to provide the most ‘functional’ information of amongst the ‘omics’ technologies as any genetic modification manifests in a change in the metabolome[44,166–168].

In this study, a targeted metabolomics approach was employed to quantify the metabolites of the phenylpropanoid pathway in three different Arabidopsis genotypes using an analytical technique based on reverse phased liquid chromatography couple with tandem

mass spectrometry. The three genotypes considered for the study were wild type (WT) plants, lines in which caffeoyl shikimate esterase (CSE) has been knocked out (*cse2*), and reduced epidermal fluorescence (ref) 8-1 lines in which the mediator complex subunits MED5a and MED5b have been disrupted (*med5a/5b ref8-1*). Both *cse2* and *med5a/5b ref8-1* lines have altered lignin content and composition (high H lignin) and superior saccharification phenotypes presenting interesting cases for targeted metabolomics. Comparative analysis was performed for each mutant relative to WT plants to depict the metabolome reorganization in the phenylpropanoid pathway as a result of the mutations.

5.3 Materials and Methods

5.3.1 Plant material

Whole stems from 5 week old Arabidopsis plants of Col-0 ecotype were harvested and quenched using liquid nitrogen. All plants included in the study were grown at a light intensity of 100 µE/m^2-s under a 16/8 hour day/night cycle in growth chambers maintained at 23^{o}C. A total of around 3 g FW of stem tissue was used for each replicate. Stem tissue was ground to a fine powder using a mortar-pestle and stored in -80^{o}C until further use.

5.3.2 Extraction of soluble metabolites

Frozen stem tissue was ground to a fine powder using a mortar and pestle to which 10 µl of the extraction solvent was added for every mg fresh weight of the harvested tissue. The extraction solvent used was 75% methanol in water with the internal standards benzoyl CoA and *p*-F-(DL)-Phenylalanine at a concentration of 1 µg/ml and 0.1 µg/ml respectively. Soluble metabolite extraction was conducted sequentially in two steps. First, the samples

were subjected to a cold extraction on a MultiTherm vortexer (Valley Park, MO) at 4^{o}C for 30 min for analysis of the labile hydroxycinnamyl CoA esters. The samples were centrifuged at 18000 g for 15 min and the supernatants (S1) were dried under a stream of nitrogen gas. The procedure was repeated for a second time by adding the same volume of extraction solvent with the only change of extraction temperature. Samples were vortexed at 65^{o}C. The supernatants (S2) collected after centrifugation were dried under a stream of nitrogen gas. Concentrations of hydroxycinnamoyl CoA esters were reported by analyzing S1 samples using LC-MS/MS, while the concentrations of the remaining intermediates of the phenylpropanoid pathway were reported as a combination of S1 and S2 samples.

5.3.3 Metabolite analysis using LC-MS/MS

Analytes were separated using a Shimadzu HPLC 20AD system on a Zorbax Eclipse C8 column (150 mm 4.6 mm, 5 μm, Agilent Technologies, Santa Clara, CA) at a column temperature of 30^{o}C and a flow rate of 1 ml/min. Metabolite detection was performed using an AbSciex QTrap 5500 triple quadrupole system equipped with an electrospray ionization (ESI) probe in the negative ion mode. Multiple reaction monitoring (MRM) was employed to monitor parent-daughter ion transitions for increased specificity and sensitivity of quantification. Chromatographic conditions and mass spectrometric parameters for hydroxycinnamoyl CoA esters and other phenylpropanoids of the pathway are as described in Chapters 2 & 3.

5.3.4 Statistical analysis

All metabolite analyses were conducted in triplicate. Data were analyzed by one-way ANOVA for independent samples using the online calculator on vassarstats.net/ (Vassar College, Poughkeepsie, NY, USA). A p-value < 0.05 was considered as a significant difference. Standard Student's *t*-test was applied to analyze differences between individual metabolite concentrations across different Arabidopsis genotypes. A modified *p*-value of 0.003 after applying the Bonferroni correction was used to establish a significant difference.

5.4 Results and Discussion

5.4.1 Profiling CSE knockout lines

Inflorescence stems from 5 weeks old Arabidopsis *cse2* lines were profiled for metabolites in the phenylpropanoid pathway. Our analysis revealed a significant change in the metabolite concentrations of *cse2* lines in comparison to wild-type Arabidopsis stems. Briefly, in *cse2* lines (i) concentrations of hydroxycinnamic acids were significantly higher in *p*-coumaric acid (~4 fold), caffeic acid (~3000 fold), ferulic acid (~16 fold), and sinapic acid(~4 fold) exhibiting several fold increase over WT pool sizes (Figure 5.1 (a)) Also we observed that, (i) caffeoyl-shikimate – substrate of CSE – accumulated over 1500 fold compared to WT (Figure 5.1 (c)), and (ii) concentrations of intermediates leading to H-lignin were significantly higher, while precursors of G and S lignin were drastically reduced (Figure 5.1 (d) & (e))

Caffeoyl-shikimate esterase (CSE) catalyzes the conversion of caffeoyl-shikimate to caffeic acid, therefore an increase in the concentration of the primary substrate of the enzyme in knockout lines (*cse2*) is expected. The *cse2* lines exhibited a reduced lignin

phenotype with significantly higher H-lignin compared to WT lines. Concurrently, we observed an accumulation in H lignin precursors specifically *p*-coumaraldehyde (~13 fold) and *p*-coumaryl alcohol (~6 fold) and a simultaneous reduction in G lignin precursors coniferaldehyde (~4 fold) and coniferyl alcohol (~12 fold) in *cse2* plants (Figure 5.1 (c), (d), & (e)). Most of our findings corroborated previous experimental observations[25]. But the most interesting and unexpected observation was a ~3000-fold increase in the caffeic acid pool when the primary route of its synthesis (CSE) is eliminated. This suggests the presence of an alternative route to caffeic acid synthesis. One possibility is the hydroxylation of *p*-coumaric acid by C3'H, but dynamic isotopic labeling studies and MFA analysis (Chapter 4) have provided evidence for another pathway to caffeic acid synthesis. Moreover, both *p*-coumaric acid and *p*-coumaroyl shikimate accumulate to over 2 fold in *cse2* lines, and the ratio of their concentrations are the same as in WT ([*p*CA]/[*p*CShK] = ~7) making it highly unlikely for the hydroxycinnamic acid (K_m > 300 µM) to compete with the shikimate ester (K_M = 7 µM) to produce caffeic acid[119]. Metabolomics data from *cse2* lines bolsters our hypothesis that there exists alternative enzyme(s) capable of hydroxylating *p*-coumaric acid to caffeic acid. Furthermore, an increase in ferulic and sinapic acids accompanied with a severe reduction in the pool sizes of their predominant precursors coniferaldehyde and sinapaldehyde respectively, indicates that caffeic acid is being converted by the action of COMT and F5H enzymes.

Effects of knocking out CSE were observed in the aromatic amino acid pathway with a reduction in Phe (Figure 5.1(a)) and tryptophan (Figure A3.1) indicating a regulatory link between the pathways. From increased pools of sinapic acid derivatives, sinapoyl-glucose and malate taken together with previous findings of increased caffeoyl and feruloyl

derivatives and lignin analysis, it can be concluded that there is a general shift of carbon flux towards H lignin and other phenylpropanoid derivatives.

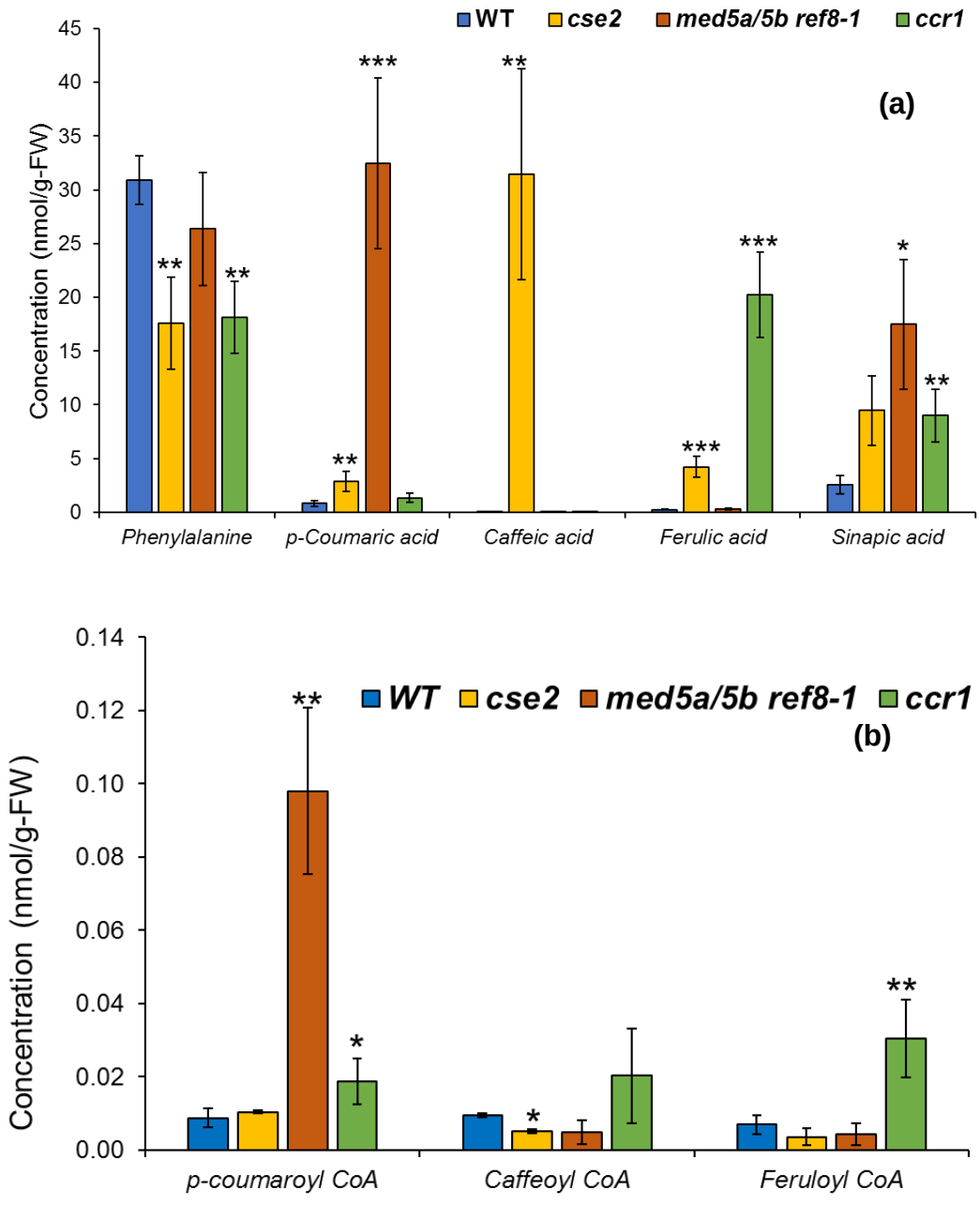


Figure 5.1: Metabolite concentrations of phenylpropanoid intermediates in 5 week old whole stems of Arabidopsis WT (blue), *cse2* (yellow), *med5a/5b ref8-1* (orange), and *ccr1* (green) lines. Data presented as mean ± S.D. from n=3 replicates. *p < 0.05, **p < 0.01, and ***p < 0.001 were obtained using standard Student's *t*-test.

Figure 5.1. continued

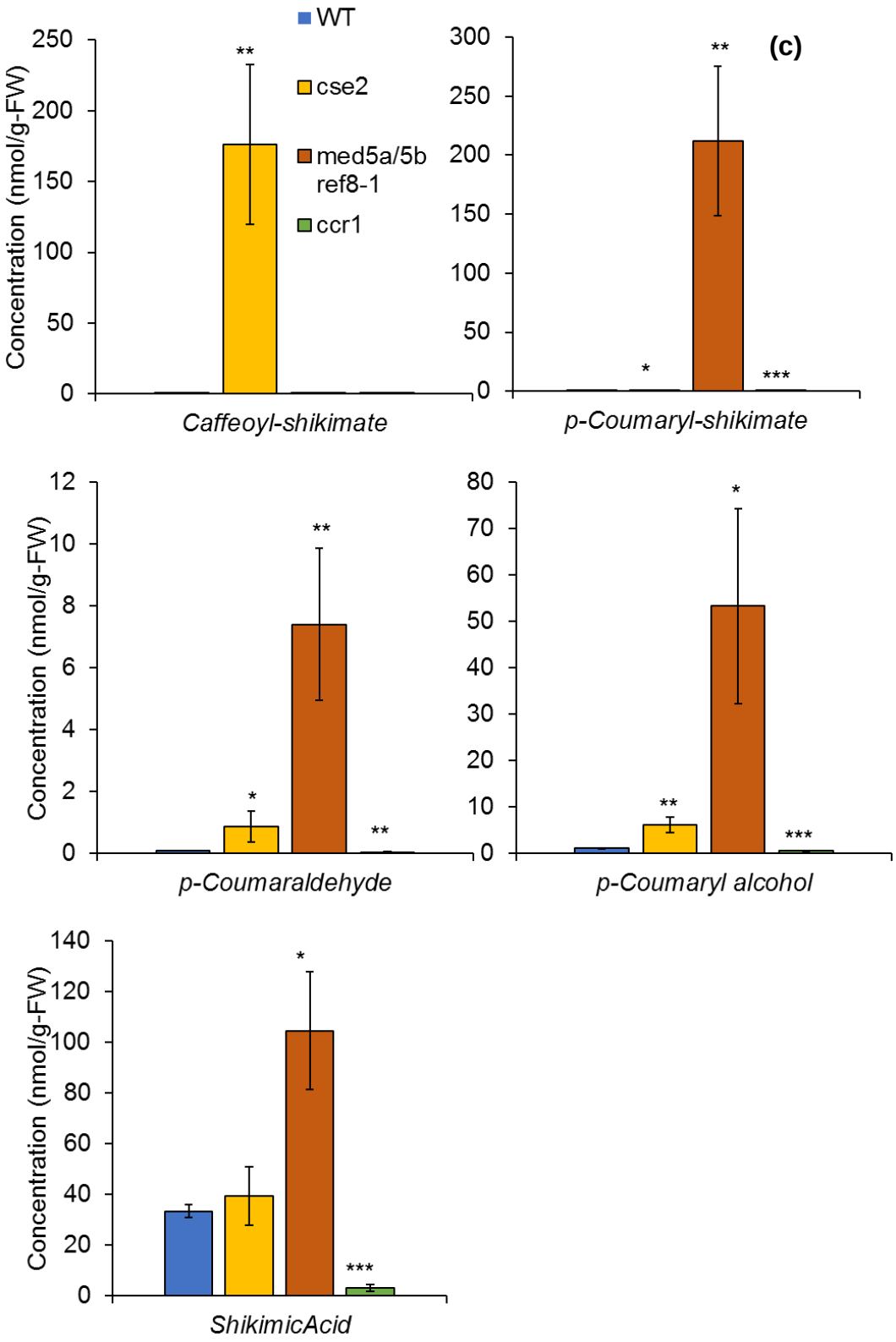
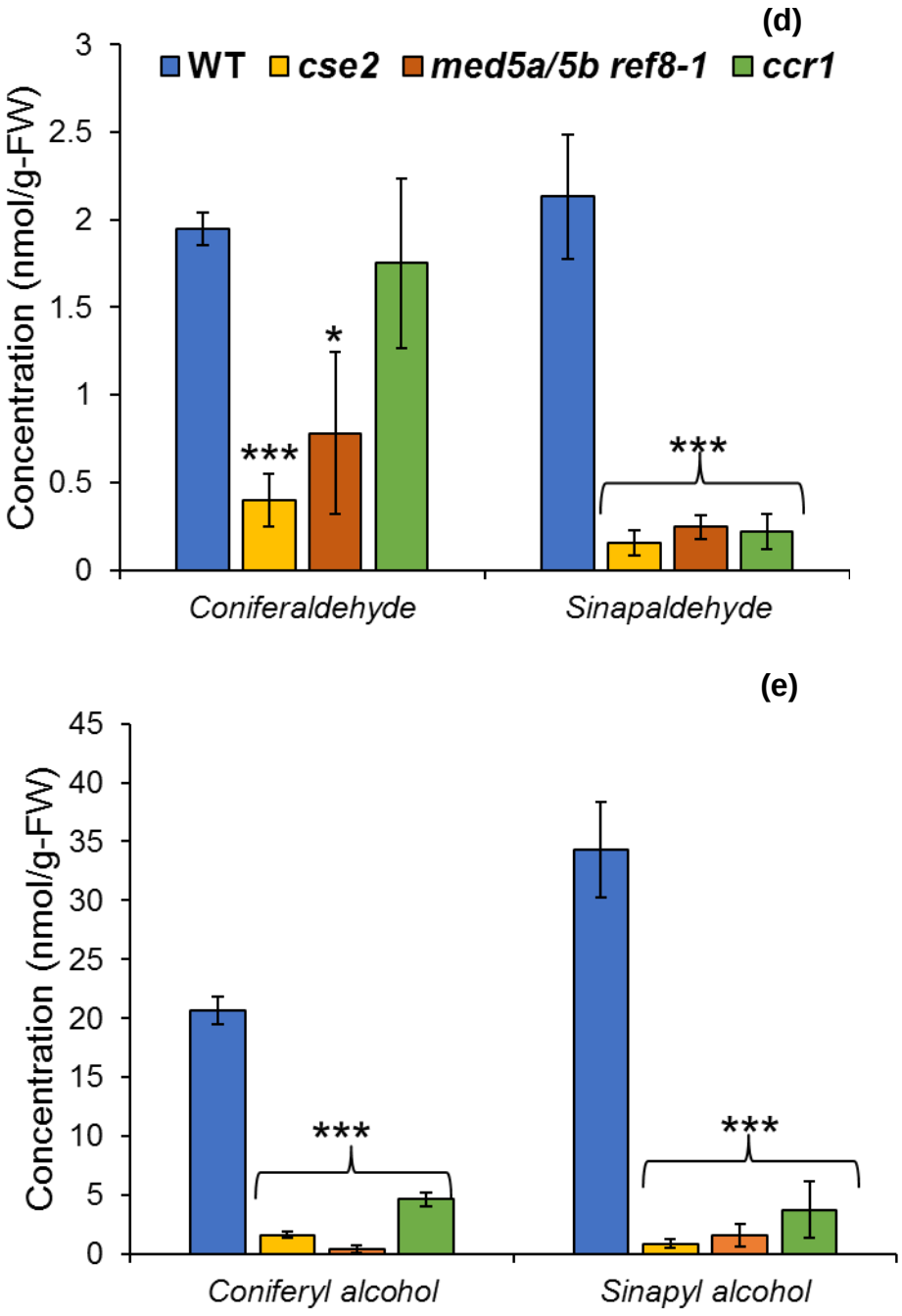


Figure 5.1. continued



5.4.2 Profiling med5a/5b *ref8-1* lines

Arabidopsis reduced epidermal fluorescence (*ref*) 8-1 lines exhibit a severely dwarfed phenotype accompanied with sterility due to a missense mutation in the gene encoding C3'H[156]. This growth phenotype is almost completely rescued on disrupting transcription regulating mediator complexes MED5a and MED5b while still retaining the high H lignin and reduced epidermal fluorescence phenotype as the *ref8-1* lines providing an interesting background for conducting targeted metabolomics of the phenylpropanoid pathway. Soluble metabolite analysis of *med5a/5b ref8-1* lines revealed (i) significant accumulation in *p*-coumaryl shikimate (~1200 fold), (ii) increased pool sizes in all the intermediates of H lignin synthesis, (iii) a significant depletion in pool sizes of the precursors of G and S lignin.

Substrate accumulation in knockdown/knockout lines of an enzyme is expected as is the case of *p*-coumaryl shikimate in *med5a/5b ref8-1* lines (Figure 5.1 (c)). In addition, a simultaneous increase in all *p*-coumaryl derivatives namely *p*-coumaric acid (~40 fold), *p*-coumaroyl CoA (~12 fold), *p*-coumaraldehyde (~90 fold), *p*-coumaryl alcohol (~50 fold) was observed in line with increased levels of H-lignin observed in the mutant lines. All metabolites downstream of C3'H such as coniferaldehyde (~2 fold), sinapaldehyde (~8 fold), coniferyl alcohol (~40 fold), and sinapyl alcohol (~20 fold) were severely reduced (Figure 5.1 (d)&(e)). Decreased metabolite pools taken together with elevated transcripts of most of the phenylpropanoid pathway genes[156], suggests that the flux towards G and S lignin is reduced.

Interestingly, higher pool sizes of caffeic, ferulic, and sinapic acid were also observed in *med5a/5b ref8-1 lines* (Figure 5.1 (a)). If C3'H were the predominant route to caffeic acid, a significant reduction in product pools would be expected. Taken together

with the increase in ferulic and sinapic acids when the corresponding hydroxycinnamyl aldehydes are significantly lower suggests an alternative synthesis route to be active as discussed previously for the *cse2* lines.

In addition, *med5a/5b ref8-1* lines exhibited a 3-fold increase in shikimic acid pools. The total concentrations of shikimate combining free shikimate, *p*-coumaryl shikimate and caffeoyl shikimate pools is ~300 nmol/g-FW, almost 10-fold higher than WT levels indicating a general upregulation of the shikimic acid pathway, which is expected in mediator disrupted lines. Although, no significant differences were seen in Phe pools, reduced levels of tryptophan and increased levels of tyrosine were observed (Figure A3.1).

5.5 Conclusions

In this study, we conducted targeted metabolomics in inflorescence stems of WT, *cse2*, and *med5a/5b ref8-1* Arabidopsis genotypes. Accumulation of precursors of H-lignin and a significant reduction in the precursors to G and S lignin suggests a general shift of flux towards lignin derived from *p*-coumaroyl sub units. Increased pools of hydroxycinnamic acids while the respective predominant precursors are highly reduced, strongly suggests an alternative route of synthesis. Although, more definitive evidence can be obtained by conducting metabolic flux analysis using isotopic labeling studies. Effects of genetic modifications in the phenylpropanoid pathway are observed in the shikimic acid and aromatic amino acid pathways. Future targeted metabolomics studies extended to include the shikimate and aromatic amino acid pathway would go a long way in elucidating the regulatory links between the metabolic networks.

6. INVESTIGATION OF SUB-CELLULAR COMPARTMENTATION USING NON-AQUEOUS FRACTIONATION

6.1 Abstract

6.2 Introduction

One of the unique features of eukaryotic cells is the compartmentalization of metabolism across several organelles. Despite the physical segregation, metabolism in every compartment, in a way, depends on other organelles of the cell for a supply of energy cofactors (ATP, NADP) or other metabolic precursors adding an additional layer of complexity and regulation to eukaryotic metabolism[44,169]. Consequently, to gain a fundamental understanding of how a eukaryotic cell functions, it is important to know how these processes are linked and connected across these compartments. Plant cells, specifically, are challenging to understand due to the presence of (i) large number of sub-cellular compartments like the plastid, vacuole, and cell walls, and (ii) more diverse metabolic networks[170].

An instance of such sub-cellular compartmentation – relevant to the current study – is the case of lignin biosynthesis *via* the phenylpropanoid pathway (Figure 6.1). Lignin engineering of feedstocks by genetically modifying enzymes of the phenylpropanoid pathway has garnered significant attention due to a focus on renewable and alternative sources of energy[6,17]. Phenylalanine, the precursor of the metabolic network undergoes deamination followed by a series of functional modifications catalyzed by 11 enzyme families to produce *p*-coumaryl, coniferyl, and sinapyl alcohols, the three fundamental units of lignin, also referred to as monolignols[171]. Rational engineering of lignin

biosynthesis necessitates a fundamental and mechanistic understanding of how carbon flux is regulated in the phenylpropanoid pathway.

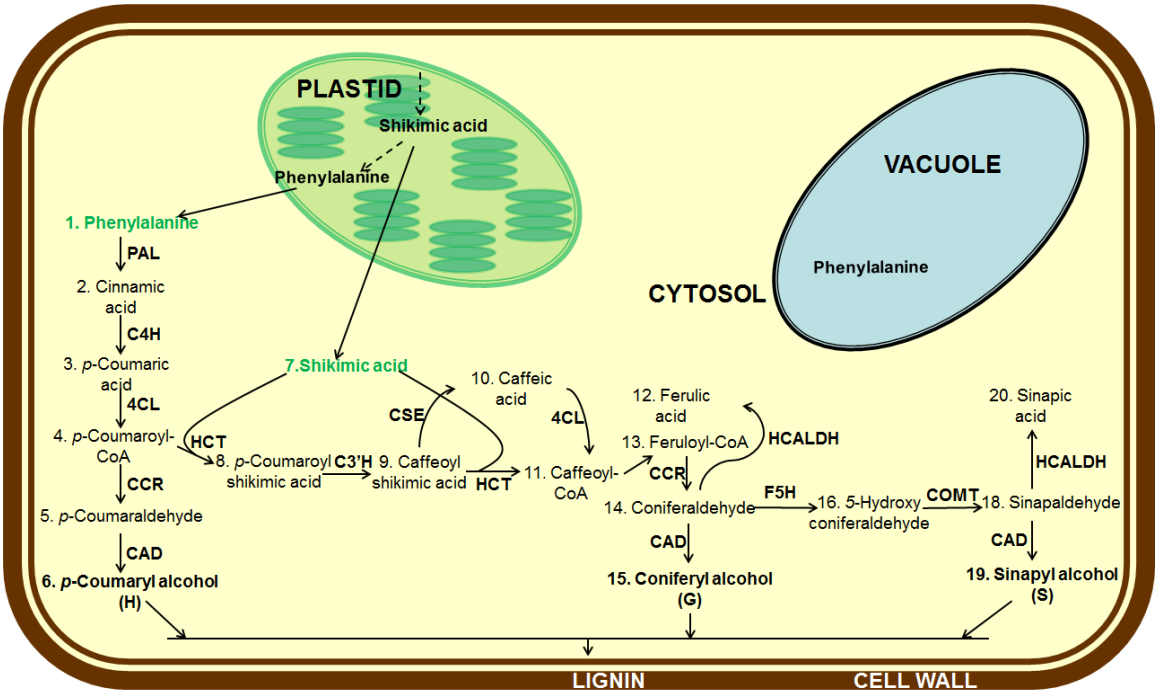


Figure 6.1: Lignin synthesis *via* the phenylpropanoid pathway. Solid arrows indicate single reactions catalyzed by enzymes as represented alongside each arrow. Dashed arrows indicate lumped reactions.

While the entire phenylpropanoid pathway is localized in the cytosol, Phe and shikimate – two metabolites participating in lignin synthesis – are first produced in the plastid and then transported to the cytosol (Figure 6.1). As a result, the relative distribution of Phe and shikimate in the plastid and cytosol may play a regulatory role in carbon allocation to lignin. Precursors to a metabolic network usually tend to have large concentration control coefficients, and Phe being the point of origin for the monolignols may exert a large control on the input flux into the pathway. Phenylalanine ammonia lyase (PAL), the first enzyme of the metabolic network and the enzyme that catalyzes the deamination of Phe is located in the cytosol. Therefore, flux into the phenylpropanoid

pathway is a strong function of the cytosolic concentration of Phe. Shikimate, in addition to being a precursor to Phe synthesis in the plastid also participates in a three reaction series catalyzed by HCT and C3'H in the cytosol (Figure 6.1). the HCT-C3'H-HCT reaction trio forms the bridging link between the H-lignin and the G and S lignin precursors[172]. HCT is a reversible enzyme that catalyzes conversion of *p*-coumaroyl CoA to *p*-coumaryl shikimate and caffeoyl shikimate to caffeoyl CoA. Shikimic acid is consumed in the former and released in the latter reaction. Given that the other P450-dependent monooxygenase enzymes of the pathway – C4H and F5H – accept unconjugated acids, alcohols and aldehydes as their substrates, it is an unanswered question as to why C3'H prefers a shikimate conjugated hydroxycinnamic acid. This led to a speculation whether shikimic acid was a putative regulatory link between phenylalanine synthesis (by the shikimate pathway) and its utilization (in the phenylpropanoid pathway). Consequently, any further investigation into the regulation of carbon flux in the metabolic network requires measurement of concentrations of Phe and shikimate in three major subcellular compartments of the cell – plastid, cytosol, and the vacuole.

Non aqueous fractionation (NAQF) has been widely used for resolving metabolite pools across different sub-cellular compartments in plants[173–179]. Unlike most cell fractionation procedures that are used to purify intact organelles, the NAQF method enriches disrupted pieces of compartments across a continuous non-aqueous density gradient. The use of non-aqueous solvents offers two main advantages over aqueous (i) quenches metabolism and prevents conversion of metabolites, and (ii) prevents reallocation of polar metabolites[178]. Removal of an aqueous environment prevents the metabolites from diffusing across the gradient, instead allowing them to co-migrate with the

compartments they were localized in. The continuous gradient is collected in 4 to 10 fractions, and marker enzymes and metabolites are measured in each fraction. Sub-cellular distribution of metabolites can be inferred by solving a system of mathematical equations representing balances on the metabolites in each fraction[180].

In this study, the NAQF technique was applied to Arabidopsis inflorescence stems to measure sub-cellular metabolite levels. Using this technique, relative distribution of Phe and shikimate pools in the cytosol, plastid and vacuole were determined.

6.3 Materials and Methods

6.3.1 Plant material

Whole stems from 5 week old Arabidopsis plants of Col-0 ecotype were harvested and quenched using liquid nitrogen. All plants included in the study were grown at a light intensity of 100 $\mu E/m^2$-s under a 16/8 hour day/night cycle in growth chambers maintained at 23^oC. A total of around 3 g FW of stem tissue was used for one gradient. Stem tissue was ground to a fine powder using a mortar-pestle and stored in -80^oC until further use.

6.3.2 Non-aqueous fractionation of Arabidopsis stem tissue

The overall procedure for NAQF using *n*-heptane and tetrachloroethylene (C_2Cl_4) as the non-aqueous solvents was adapted from previously published methods on Arabidopsis leaves (Figure 6.2). The following sections provide a basic description of the entire method. A list of more detailed protocols and procedures, optimized and altered for Arabidopsis stem tissue, have been documented in the appendix ().

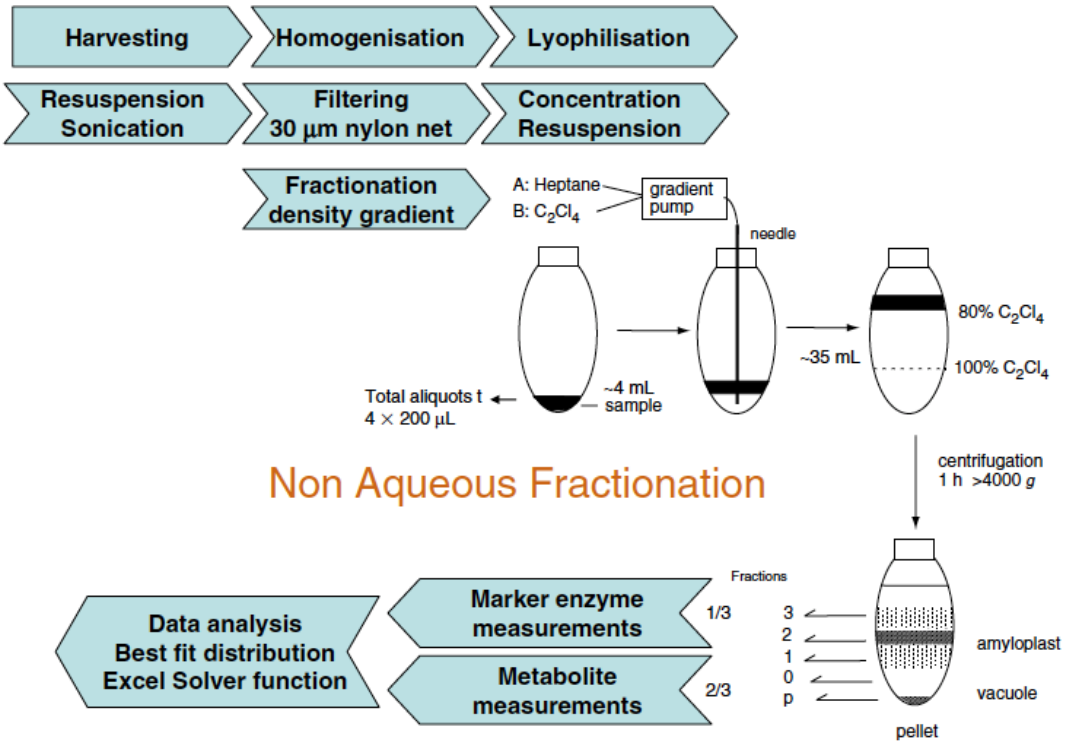


Figure 6.2: Schematic of overall procedure of NAQF adopted from Geingenberger *et al*, 2011.

Tissue lyophilization and homogenization. Frozen and pulverized Arabidopsis stem material was lyophilized in a Labconco freeze dryer over a period of 3 days in 45 ml Eppendorf tubes. Each tube had material enough to occupy only the conical section of the Eppendorf tube to prevent formation of a cake at the onset of freeze drying and impede sublimation of ice. Tubes were slightly inclined in the glass jars to increase surface area for efficient drying. After lyophilization, the dried tissue was placed in a dessicator with drierite until further use. A known amount of dried stem tissue (~300 mg-DW) was further homogenized in a Retsch ball mill in 6 ml of 34:66 %v/v solution (ρ = 1.33) of n-heptane (d = 0.684) and C_2Cl_4 (ρ = 1.62) in two cycles of 10 mins at a frequency of 30 s^{-1}. The bead beating chambers were half filled with 1 mm stainless steel beads. Homogenization by bead beating was followed by ultrasonication using a Misonix XL-2000 sonicator

(Farmingdale, NY) with a CML-4 probe at setting 13 in 6 cycles of 15 s pulses followed by a 10 s rest between cycles. Sample was placed on ice during the rest cycle to prevent thermal damage of enzyme. The homogenized material was filtered through a nylon cloth (<22 µm) and centrifuged at 4000x*g* for 15 min at 4^{o}C. The pellet was re-suspended in 5 ml of 1.33 g.cm^{-3} solution of *n*-heptane and C_2Cl_4 to be deposited onto the gradient.

Sample separation using a continuous non-aqueous density gradient. A 35 ml linear density gradient from 1.3 – 1.62 d.cm^{-3} was layered in a 50 ml centrifuge tube (Table A5.1). The re-suspended sample was inserted on top of the gradient and centrifuged using a Labconco floor centrifuge at 4^{o}C for 90 min at 13000 rpm. Most of the sample material was focused in the top and middle fractions while the heaviest fraction of the biomass formed a pellet at the bottom of the centrifuge tube(Figure 6.3). The gradient was divided into 6 fractions (F1 to F6) of 5-7 ml each and the pellet re-suspended in 3 ml of C_2Cl_4 constituted the seventh fraction (F7). Each fraction was further divided into three sub-fractions for enzyme assays (F1-E1, F2-E2 etc.) and metabolite analyses (F1-M, F2-M, etc.) in the ratio 1:1:2.

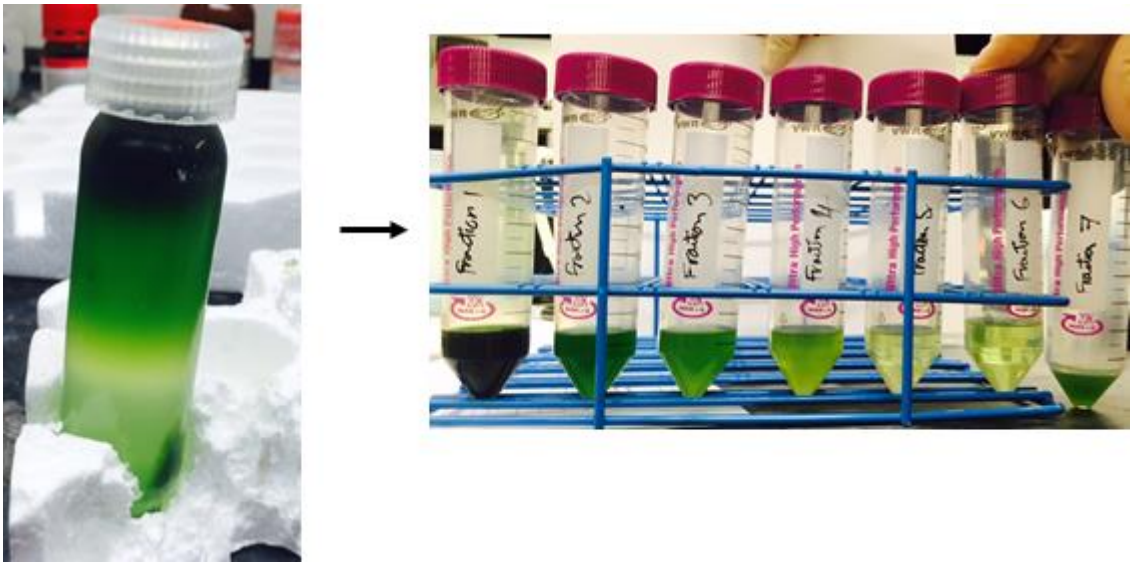


Figure 6.3: Density gradient with homogenized Arabidopsis stem tissue after centrifugation. The entire gradient was divided into 6 fractions (F1-F6) and the pellet was resuspended in C_2Cl_4 for the seventh fraction (F7)

Marker enzyme assays. Phosphoenolpyruvate carboxylase (PEP-C; E.C. 4.1.1.31), ADP-glucose pyrophosphorylase (AGPase; E.C. 2.7.7.27), and α-mannosidase (E.C. 3.2.1.24) were chosen as marker enzymes for the cytosol, plastid, and vacuole respectively. All buffers for assays and protein extraction were prepared according to the procedures detailed in the Appendix B1 & B2. The sub-fractions separated for enzyme assays (F1-E1, F2-E2, etc.) were dried under a stream of nitrogen gas until complete removal of non-aqueous solvent. The dried pellet was then suspended in 0.5 ml of enzyme extraction buffer by vortexing in a MultiTherm vortexer at 4^{o}C for 15 min. The suspensions were centrifuged in a microcentrifuge equilibrated at 4^{o}C. The supernatant was decanted and refrigerated until further use in the enzyme assays.

(i) ***Cytosolic PEPc assay****:* The master mix(MM) for the assay was made of stocks solution of 110 mM Tris sulfate, adjusted to a pH 8.5 using NaOH, 300 mM $MgSO_4 \cdot 7H_2O$, 6 mM β-NADH, 100 mM $NaHCO_3$, 1,4-dioxane, 300 mM

dithioerythritol (DTE), 600 U/ml malic dehydrogenase (MDH). The volumes and final concentrations of each component are detailed in the appendix. Phospho*enol*pyruvate (PEP) at a concentration of 30 mM was used as the substrate (S).

To measure PEPc activity, 10 µl of the protein extracted from each of the 7 fractions was aliquoted in triplicate and added to 280 µl of MM in wells of a 96-well plate. Following this, 10 µl of the S solution was added to each well except the blank. The plate was mixed thoroughly for 1 min after the addition of the substrate. Enzyme activity was determined by measuring the disappearance of β-NADH at a wavelength of 340 nm and a temperature of 25^{o}C over a period of 10 minutes using a Molecular Devices SpectraMax UV-Vis spectrophotometer.

(ii) ***Plastidial AGPase assay***: The master mix(MM) for the assay was made of stocks solution of 100 mM HEPES $MgCl_2$ buffer, adjusted to a pH 8 using NaOH, 300 mM phosphoglyceric acid (PGA), 300 mM dithiothreitol (DTT), 10 mM β-NADP, 10 mM ADP glucose, 1 mM of glucose-1,6-diphosphate, 885 U/ml of phosphoglucomutase (PGM) from rabbit muscle, 250 U/ml of glucose-6-phosphate dehydrogenase. The volumes and final concentrations of each component are detailed in the appendix. Sodium pyrophosphate ($NAPP_i$) at a concentration of 25 mM was used as the substrate (S).

To measure AGPase activity, 50 µl of the extracted enzyme from every fraction was aliquoted in triplicate and added to 200 µl of the MM in wells of a 96-well plate. Following this, 30 µl of the substrate was added to each well except for the blank. The plate was mixed thoroughly for 1 minute after the addition of the

substrate. Enzyme activity was determined by measuring the appearance of β-NADPH at a wavelength of 340 nm and a temperature of 25ºC over a period of 10 minutes using a UV-Vis spectrophotometer.

(iii) Vacuolar α-mannosidase assay: Citrate buffer (CB) at a concentration of 100 mM was adjusted to a pH of 4.5 using NaOH. Borate buffer at a concentration of 200 mM and pH adjusted to 9.8 by NaOH was used as a stopping buffer. 20 mM *p*-nitrophenyl-α-D-mannopyranoside was used as the substrate for the vacuolar assay.

For the assay, 10 µl from each fraction was added to 48 µl of the citrate buffer in triplicate in wells of 96 well plate. Following this, 48 µl of the S was added to each well except for the blank, and the microplate was incubated at 37ºC for 30 minutes. The reaction was quenched by adding 194 µl of borate buffer. Enzyme activity was determined by measuring the absorbance at a wavelength of 405 nm and a temperature of 25ºC.

Metabolite analysis. All seven sub-fractions separated for metabolite analysis (F1-M, F2-M etc.) were dried under a stream of nitrogen until complete removal of the non-aqueous solvents. To the dried pellet, 500 µl of 75% (v/v) methanol in water was added and vortexed on a Multitherm vortexer at a temperature of 65ºC. The samples were centrifuged on a Beckman Coulter microcentrifuge at 15000x*g* for 15 min, after which 10 µl of the supernatant was injected on the LC-MS for analysis. Phe and shikimic acid were quantified using a previously published analytical method[102]. Metabolite abundances (number of moles in one fractions) in each fraction was normalized to the total metabolite abundance (total number of moles from all seven fractions) obtained from all the fraction.

Estimation of relative distribution of metabolites. Sub-cellular levels of metabolites were estimated by solving a system of linear equations that represent mass balances on the metabolites in each fraction as shown in Equation 6.1.

$$f_{M_p}.f_{p_i} + f_{M_c}.f_{c_i} + f_{M_v}.f_{v_i} = f_{M_i} \qquad i=1 \text{ to } 7;\ \text{Equation } 6.1$$

In the above equation, f_{M_p}, f_{M_c}, and f_{M_v} represent the fraction of the metabolite (M) in the plastid, cytosol, and the vacuole respectively. These variables – also the unknowns in our study – add up to one, due to the assumption that the metabolite is localized only in the three compartments considered in the study. The fraction of the plastid, cytosol, and the vacuole in fraction *i* are denoted by f_{p_i}, f_{c_i}, and f_{v_i}, the values for which are obtained from the marker enzymes assays. The term on the right had side of the equation represents the fraction of metabolite in fraction *i*, the values for which are obtained from analyzing metabolite concentrations using LC-MS. Such an equation can be written for every fraction (F1 to F7) resulting in a system of 7 linear equations that were simultaneously solved to obtain the relative distributions of metabolites.

Error propagation analysis. Means and standard deviations for the fractions of compartments were calculated using data measured in triplicates. Using the means and standard deviations, 1000 synthetic data sets representing the relative abundances of compartments in each fraction were generated using a normal distribution sampler (*normrnd*) in MATLAB. The system of linear equations was solved for each data set to obtain a set of 1000 solutions. The relative distribution of metabolites in the three

compartments and the error of estimation were reported as mean and standard deviations of the 1000 solutions.

6.4 Results and Discussion

6.4.1 Distribution of sub-cellular compartments across the gradient.

Particles settle in different layers of the gradient as a function of their density thereby resulting in a partial enrichment of enzymes and metabolites of a sub-cellular compartment across the entire gradient. Less dense compartments, accompanied by the associated enzymes and metabolites, tend to settle in the top fractions of the gradient, while the denser compartments are increasingly enriched in the bottom fractions or the pellet fraction[180]. The relative abundance of a sub-cellular compartment in a fraction is obtained from normalized marker enzyme activities as previously described (Materials and Methods). Our analysis of marker enzyme assays on Arabidopsis stems subjected to NAQF, indicated a higher enrichment of plastid in the lighter (top) fractions and the vacuolar material in the heavier (bottom) fractions and the pellet (Figure 6.4). No significant trend was observed in case of the cytosolic marker because (i) the inherent error of measurement of the assay was relatively high, (ii) the cytosol is associated with both plastidial and vacuolar membranes in the cell thus can co-settle with either compartment (Figure 6.4). These findings are in accordance to previous studies in Arabidopsis leaves [175], barley seeds [173,181], spinach leaves [179], rose petals [176] which consistently placed the plastid in the lighter and vacuole in the heavier fractions of the gradient.

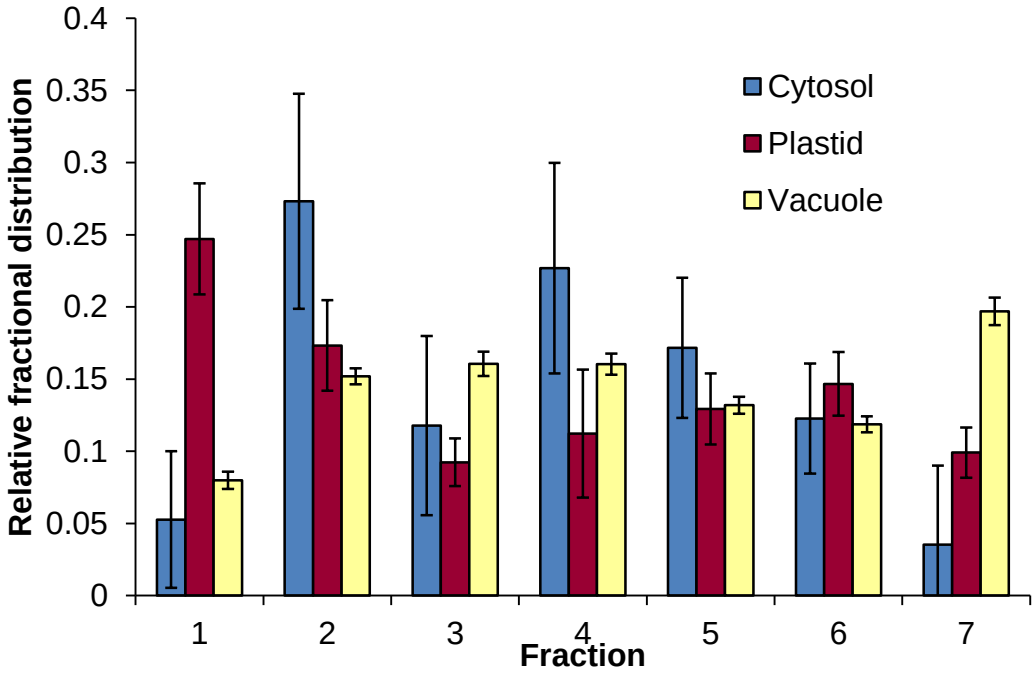


Figure 6.4: Relative distribution of the plastidial (magenta), cytosolic (blue), and the vacuolar (yellow) compartments across the density gradient. Numbers on the x-axis represent fractions with 1 being the lightest and 7 being the pellet or the heaviest fraction. Data presented as mean ± S.D from n=3 replicates.

6.4.2 Relative sub-cellular distribution of Phe and shikimate.

Compartment enrichments and metabolite abundances in each fraction were used to solve for the relative sub-cellular distribution of metabolites. The majority of Phe was estimated to be localized in the cytosol and around 34% of the total Phe pool measured was found to be in the plastid (Table 6.1). Although, NAQF was never previously used on Arabidopsis stems, a similar compartmental distribution of Phe was observed in Arabidopsis leaves[175]. In the case of shikimic acid, almost an even cytosolic and plastidial distribution was observed with 40% of the total pool localized in each compartment. Although the cytosol seems to be the dominant compartment, the local concentration of the metabolites in the plastid is higher due to lower organelle volume maintaining a concentration gradient for transport to the cytosol. Vacuolar pools are for

both metabolites are around 15-18% of the total pool, but are accompanied with large errors of estimation making them statistically insignificant (Table 6.1). However, a recent study has shown that Phe hyperaccumulating plants sequester Phe into the vacuolar compartment (Lynch et al., 2017; submitted). In other words, even if significant pools of Phe may not be localized in the vacuole in wild-type plants, including the vacuolar compartment in future NAQF studies on Arabidopsis is imperative as such a vacuolar sequestration is possible in transgenic lines.

Table 6.1: Partitioning of metabolites across different sub-cellular compartments.

Metabolite	Cytosol (%)	Plastid (%)	Vacuole (%)
Phenylalanine	51.1 ± 28.5	34.1 ± 22.3	14.8 ± 18.9
Shikimic acid	40.8 ± 21.2	40.9 ± 16.9	18.3 ± 16.8

Confidence in relative distribution estimates of metabolites was evaluated by propagating errors of measurement from marker enzyme assays as described in the Materials and Methods section. The large deviations in estimated values of sub-cellular distribution of metabolites arises from the large deviations observed in cytosolic marker assays, specifically in fractions 1 and 7 most likely due to low amounts of cytosolic material (Figure 6.4). One solution for this would be to increase the amount of biomass that is inserted on the gradient to ensure significant biomass distribution across different fractions to obtain absorbances above limits of detection while performing enzyme assays. Another solution maybe to reduce the number of fractions into which the gradient is divided. Although this may ameliorate variances in enzyme assay measurements, it would reduce the degrees of freedom in estimating the compartmental distribution of metabolites[182]. There have been studies where NAQF was optimized for the number of fractions based on

the standard deviations of assay measurements and estimates of compartmental distribution, where between 6-10 fractions were suggested[180,182,183]. However, the optimal number of fractions would be tissue specific (leaf, stem, roots etc.) and should be evaluated anew when working with different systems.

6.5 Conclusions

Knowledge of distribution of metabolite pools across different compartments is advantageous in designing rational metabolic engineering experiments, specifically when the metabolites involved may have a regulatory role as in the case of Phe and shikimate in lignin biosynthesis. We employed non-aqueous fractionation to Arabidopsis stems to estimate the relative sub-cellular distribution of Phe and shikimate. Given their synthesis in the plastid, it was expected to observe the metabolites localized in the plastid and the cytosol, although no significant vacuolar pools were estimated. Smaller sub-cellular volume of the plastid would result in a higher local concentration of metabolites compared to the cytosol, ensuring a gradient across the compartments for transport. Some applications for this technique could be (i) to estimate distribution – specifically in the vacuole – in transgenic lines that see significant accumulation of phenylpropanoid intermediates, (ii) estimate metabolite distribution in Arabidopsis stems fed with labeled precursors (such as $^{13}C_6$-Phe) when estimating fluxes or developing a kinetic model for the phenylpropanoid pathway, (iii) to include metabolites of the shikimate and aromatic amino acid pathway in the plastid to gain insight into how they link with the phenylpropanoid pathway.

7. MACHINE LEARNING DRIVEN ESTIMATION OF AN OPTIMAL LIGNIN PHENOTYPE IN ARABIDOPSIS FOR IMPROVED SACCHARIFICATION

7.1 Abstract

Recalcitrance of lignocellulosic biomass to saccharification is a major impediment to the economical production of biofuel. Consequently, the past two decades have witnessed several genetic engineering efforts targeting lignin biosynthesis in bioenergy crops to improve saccharification yields. Although several of these studies found an overall negative correlation between lignin content and saccharification efficiency, the relationship between the composition of lignin on sugar extractability is complex and not well understood. In this study, we implemented support vector machine (SVM) based regression to predict the saccharification efficiency and biomass yields (plant height) of *Arabidopsis thaliana* plants as a function of total lignin content, and the composition of the monomers that make up lignin, namely *p*-coumaryl (H), coniferyl (G), and sinapyl alcohol (S) derived lignin. The model was developed and validated on data acquired from 9 independent studies totaling 53 Arabidopsis lines encompassing several genotypes. Data was artificially generated using standard deviations reported in literature for the input and output variables in order to serve two purposes (i) generate a considerable data set for obtaining higher regression performance, and (ii) obtain a more even distribution of the input variables to ameliorate prediction bias. A total of 500 data sets were sampled using the empirical bootstrap technique. Predictions from the trained and cross-validated SVM models resulted in an acceptable agreement to experimental data for both saccharification efficiency (R^2~0.92) and plant height (R^2~0.73) on validation data sets. The SVR models also

successfully predicted the saccharification efficiency and plant height of Arabidopsis transgenic lines that were not included in the training data.

In addition, functional forms obtained as a result of SVM regression were optimized using genetic algorithms (GA) to predict the optimal lignin content and composition that maximizes the product of saccharification efficiency and plant height, which is representative of the total sugar yield for conversion to biofuel. This effort produced two optimal solutions that both indicated a moderately lower lignin content to be conducive to sugar extractability, but interestingly with varying H:G:S composition.

7.2 Introduction

Biofuel production utilizing sugars localized in the secondary cell walls of lignocellulosic feedstock is a potential and sustainable alternative to fossil fuels as a source of energy[2,3,184]. Localized in the plant cell walls, lignin has long been known as a major contributor to biomass recalcitrance as it entraps useful cell wall polysaccharides rendering them inaccessible for enzymatic hydrolysis – a process called saccharification – thereby making biofuel production a highly cost intensive process[6,7,26]. Lignin is primarily made from *p*-hydroxyphenyl (H), guaiacyl (G), and syringyl (S) units that are derived from the products of the phenylpropanoid pathway – a metabolic network that is well established and conserved in all vascular plants – namely *p*-coumaryl, coniferyl, and sinapyl alcohols[20,110].

As a result, the past few decades have witnessed several genetic engineering efforts targeting lignin biosynthesis in various plant systems[8,15,17]. Although these efforts demonstrated an overall negative correlation of lignin with saccharification efficiency, the effect of lignin composition on saccharification efficiency is not well understood. Early

studies have suggested a high S/G ratio is conducive for improved saccharification[14,33], but there have been reports of transgenic lines that had a lower or an unchanged saccharification phenotype compared to wild type plants in spite of having a high S/G ratio[16,32]. In some cases, lines that had H lignin units exhibited superior saccharification efficiencies[25,185]. Furthermore, genetic engineering experiments often result in pleiotropic effects. For example, a line with reduced lignin can have an altered lignin composition, structural defects in plant vasculature resulting in growth defects and dwarfism, changes in cell wall structures and accumulation of cellulosic and hemi-cellulosic sugars.

In order to harness the above repository of information for rational engineering of biofuel crops with improved cell wall characteristics and higher biomass yields, a multivariate approach in understanding the highly non-linear relation between the biological traits and the target phenotypes is necessary. There have been previous studies that analyzed the relation between biological traits and saccharification[35,186–188]. Although detailed, these studies largely stressed on one to one correlations between variables or attempted to linearly map various structural and biological traits to digestibility.

In the context of the aforementioned, the objectives of this study were: (i) to evaluate the empirical functional forms that relate total lignin content and composition to saccharification and growth phenotypes respectively using support vector machine (SVM) regression; (ii) to test and validate the regression model(s) on Arabidopsis mutant lines not included in training; (iii) to estimate optimal lignin content and composition that maximizes the total saccharification yield using genetic algorithms (GA).

Support vector machines (SVM) is a supervised machine learning algorithm that is generally used as a pattern recognition tool and a binary classifier, but can be modified for use as a regression technique[189,190]. SVM is well known for its ability to model non-linear relationships and employs a quadratic programming problem based regression function, the solution to which is global and generally unique. SVM in combination with optimization techniques such as GA, finds applications in a wide array of disciplines[191]. In this study, SVM based regression technique was used to obtain functional relationships between the explanatory variables (total lignin content, %H, %G, and %S lignin composition) and the response variables (%saccharification efficiency and plant height) in *Arabidopsis thaliana*. Data from 53 Arabidopsis lines were collected from the literature across 9 independent studies (Table 7.1). Training data sets – after data augmentation and pre-processing – were generated using empirical bootstrap sampling. The trained SVM models were successfully validated on mutant lines that were not included in the training data. The SVM models were further optimized using genetic algorithms (GA) to obtain the optimal values of the input variables that maximized the total saccharification yield.

Table 7.1: List of Arabidopsis plants considered for the study

S. No	Lines	Reference
1.	***WT Col-0, pal1-2, pal1-2, pal2-2, pal2-3, 4cl1-1, 4cl1-2, 4cl2-1, 4cl2-3,ccoaomt1-3, ccoaomt1-5, ccr1-3, ccr1-6, f5h1-2, f5h1-4, comt-1, comt-4, cad6-1, cad6-4***	[35]
2.	***WT Col-0, cse-1, cse-2***	[25]
3.	***WT Col-0, cse-2, cse-2 proVND7::CSE#1, cse-2 proVND7::CSE#2, cse-2 proVND7::CSE#3, cse-2 proVND7::CSE#4***	[185]
4.	***WT Col-0, C4H-F5H, med5a/5b ref8-1, fah1-2, COMT1***	[192]
5.	***WT Col-0, C4H::qsuB-1, C4H::qsuB-3, C4H::qsuB-6, C4H::qsuB-7***	[63]
6.	***WT Col-0, med5a/5b, med5a/5b ref8-1***	[156]
7.	***WT Col-0, lac4-2, lac17, lac4-1 lac17, lac4-2 lac17***	[193]
8.	***WT Col-0, ubiC-pobA-1, ubiC-pobA-2, ubiC-pobA-3***	[194]
9.	***WT Col-0, fpgs1-1***	[132]

7.3 Materials and Methods

7.3.1 Data Collection and Processing

The database for the study consisted a total of 53 Arabidopsis lines – including wild-type and transgenics – from 9 independent studies (D_0, Table 7.1) that reported the variables of interest to this study. AcBr lignin (% CWR), lignin composition (%H, %G, %S lignin) constitute the input variables, and plant height (cm) and saccharification efficiency (% cellulose) from untreated biomass constitute the output variables. All instances of

unreported height for wild-type lines were valuated as an average of heights corresponding to wild-type lines in the remaining studies. Plant heights for transgenic lines that showed no growth phenotype were assigned an average height equal to the wild-type lines. In cases where saccharification efficiencies have been reported in different units of measurement, they have been either converted to the default units of % cellulose using information in the study, or have been scaled using wild-type lines from other studies as a reference.

7.3.2 Data Augmentation

The database of 53 lines was augmented by randomly generating data assuming a normal distribution on the input variables. Standard deviations reported from the experimental measurements in the references were used for this procedure. For all the lines where plant height was not reported, the standard deviations were calculated using wild-type lines from all the remaining studies. The lines corresponding to high %H and high %S lignin (Figure A4.1) were underrepresented making up only 12 data points out of the total of 55. In order to make the distribution of the input variables more uniform, 30 data points were generated for each underrepresented line and 15 for the other 43 lines bringing the total size of the database to 1060 data points. Of these, a subset of 335 data points was randomly selected for model validation. The remaining 725 data points were used for training and testing the SVR model.

7.3.3 Support Vector Regression

SVM is a machine learning technique developed by Vapnik & co-workers in 1997 that largely finds application in binary classification, patter recognition and regression. SVM

implements structural risk minimization (SRM) inductive principle that allows it to achieve a generalized model by balancing the quality of fit to the training data against the complexity of the model. Support Vector Regression (SVR) is a version of SVM proposed by Vapnik et al. in 1997[195] as a nonparametric regression technique thereby obviating an *a priori* knowledge of the analytical relation between the input and target variables.

Given a typical training dataset $D = \{(x_i,y_i)\}^n \in R^d \times R$, where x_i and y_i are the i^{th} pair of input and output variables respectively of a total set of n data points, SVR aims to find a function $f(x)$ that has a deviation of no more than ε from the target value of y_i for the entire training data set. SVR achieves this by implementing the following estimation function:

$$f(\mathbf{x}) = \mathbf{w} \times \Phi(\mathbf{x}) + b,\ \Phi: R^n \rightarrow H,\ w \in H, \tag{1}$$

where **w** and b are coefficient of regression, $\Phi(x)$ represents the high-dimensional feature space that the input space is non-linearly mapped using a kernel function:

$$K(x_i,x_j) = exp(-\gamma \left\| x_i - x_j \right\|^2) \tag{2}$$

where γ is the kernel parameter that governs the width of the Gaussian function. The coefficients of regression are estimated by optimizing the regularized risk function:

$$\text{minimize: } \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \tag{3}$$

$$\text{subject to: } \begin{cases} y_i - \mathbf{w}.\Phi(x_i) - b \leq \varepsilon + \xi_i \\ -y_i + \mathbf{w}.\Phi(x_i) + b \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \quad (4)$$

where $\|\mathbf{w}\|^2/2$ represents the Euclidean norm included in the objective function to ensure flatness of the function and avoid over-fitting. The slack variables ξ_i, ξ_i^* denote the distance between the actual values and the ε deviation. Parameter *C*, known as the box constraint or the cost parameter, governs the balance between tolerance for training errors and model generalizability. Together *C*, ε, and γ constitute the set of hyper-parameters that can either be user defined or optimized for further improvement in prediction performance of SVR. The default values of *C, ε, and* γ are iqr(Y)/1.349, iqr(Y)/13.49, and 1, respectively; where iqr(Y) denotes the interquartile range of the target variable (%saccharification efficiency or plant height).

In this study, SVR was applied to obtain the functional forms $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ that correlate saccharification efficiency and height to the input variables (%AcBr lignin, %H, %G, %S lignin) respectively. The radial basis function (RBF), also known as the Gaussian kernel was used for this study. All simulations were performed using the *fitrsvm* function in the Statistics and Machine Learning Toolbox offered in MATLAB R2017a package. The hyper-parameters were optimized using the *bayesopt* solver to minimize the *k*-fold cross-validation loss on the training data and improve SVR prediction performance.

7.3.4 Genetic Algorithms

GA are a widely-used technique of optimization that is inspired by the Darwinian principle of natural selection[196]. Employing a probabilistic searching method, GA creates a population of individual solutions and repeatedly modifies them to produce a new generation of solutions using specific rules for selection, crossover, and mutation[197,198]. With each generation, the population of solutions 'evolves' towards an optimal solution. GA has been known for its versatility in application, specifically when dealing with highly non-linear objective functions.

In this study, GA was used to optimize both saccharification efficiency and biomass yields (plant height) to obtain the highest net yields of sugar. This was achieved by formulating an objective function, alternatively known as fitness function, that is a product of SVM models representing saccharification efficiency and plant height for a given input **x**:

maximize: $f_1(\mathbf{x}) * f_2(\mathbf{x})$

subject to: $\begin{cases} x_2 + x_3 + x_4 = 100 \\ 6 \leq x_1 \leq 35 \\ 0 \leq x_1, x_2, x_3 \leq 100 \end{cases}$

where x_1, x_2, x_3, x_4 denote the input variables %AcBr lignin, %H, %G, %S lignin respectively. The upper bound of 35% for lignin was chosen based on the maximum value reported for total lignin per dry weight of biomass[199]. Optimization was performed using the *ga* function under the Global Optimization Toolbox in MATLAB 2017a. Adaptive

mutation and intermediate crossover functions were used for their applicability to linearly constrained systems.

7.3.5 Empirical Bootstrap Sampling

Popularized by Bradley Efron, the empirical bootstrap is one of many resampling methods that has been widely used in estimating properties of a statistic[200,201]. The fundamental idea behind empirical bootstrapping is to draw a large number of samples from a single original dataset with or without replacement. The statistic of interest is computed on each sample leading to a distribution that is close to the true population distribution, and can hence be used to estimate confidence intervals on that statistic.

In this study, empirical bootstrap sampling was employed on the training data set where 500 samples were randomly drawn with replacement from an original training data of 725 data points (D_1,D_2,..,D_{500}). Each bootstrap was of the same size of the original training data set. This was achieved by using the *datasample* function under the Statistics and Machine Learning Toolbox in MATLAB 2017a. To ensure a more uniform distribution in every randomly drawn sample, lines with a saccharification efficiency higher than 40% and plant heights less than 30 cm were weighted 10-fold higher than the remaining data.

7.3.6 Overall SVR-GA Methodology

The 500 data sets – consisting of 725 data points each – were formed by using empirical bootstrapping and trained using SVR resulting in 500 corresponding models for predicting %saccharification efficiency and plant height. Each concomitant model pair trained on a single data set was then used in the objective function formulated for GA

optimization (Figure 7.1). Each optimization cycle outputted a plausible global optimum resulting in a total of 500 solutions. The optimal solution(s) is reported as the mean of the distribution obtained as a result of this effort.

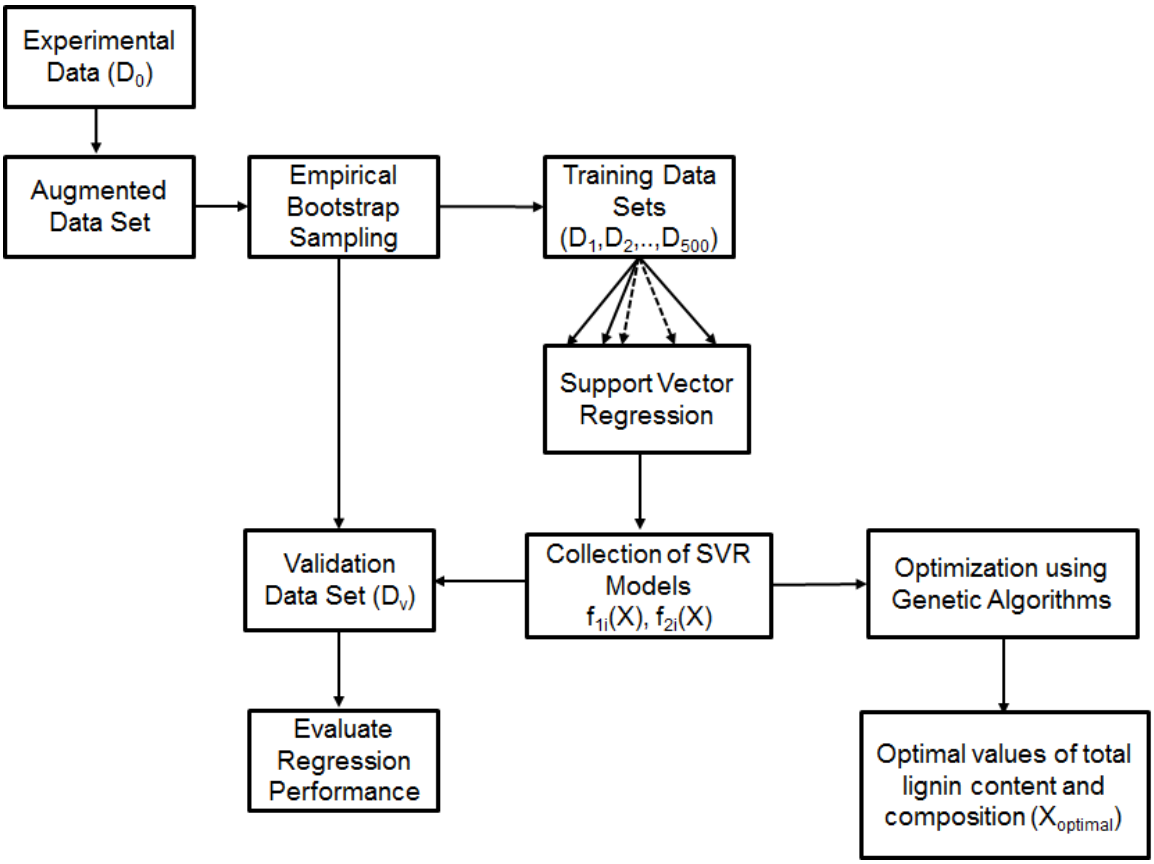


Figure 7.1: Overall framework of the SVR-GA methodology to estimate the optimal lignin content and composition that maximizes the net saccharification yield.

7.3.7 Plant Material

Arabidopsis (*Arabidopsis thaliana*) triple mutants *4cl1 4cl2 4cl3* and *4cl1 4cl3 4cl4* plants were grown in growth chambers at 22°C under 16/8 hour day/night conditions and light intensity of 120 µE m^{-2} s^{-1}.

7.3.8 Total Lignin Analysis

Cell wall residue was isolated and prepared as described previously[139]. Mature stems of Col-0 wild type, *pal1pal2pal3pal4* and *4cl14cl24cl3* mutants were harvested and pulverized in liquid nitrogen. After addition of 30 mL of 50 mM NaCl, the pulverized tissue was refrigerated overnight at 4^{o}C then centrifuged for 10 min at 4000 rpm. The pellet was then extracted with 80% ethanol and vortexed for 15 min at 65^{o}C. The extraction procedure was repeated five times. The same procedure was repeated once using acetone as the solvent. Total lignin content was measured using the acetyl bromide-soluble lignin method as described previously[140,141]. An extinction coefficient of 17.2 was used to calculate the acetyl bromide-soluble lignin[139].

7.3.9 Lignin Composition Analysis by DFRC

Lignin composition was analyzed by performing DFRC analysis as previously described[142]. Briefly, the samples prepared for acetyl bromide-soluble lignin analysis were dried down using a nitrogen concentrator and dissolved in a solvent containing dioxane/acetic acid/water (50/40/10, %v/v/v). This mixture was then reacted with Zinc dust and the products acetylated with pyridine/acetic anhydride mixture (40/60, %v/v). The acetylated lignin derivatives were quantified using gas chromatography-mass spectrometry using standard calibration curves after accounting for the response factors from the internal standard.

7.3.10 Saccharification Assays

The structural carbohydrates of untreated biomass were determined using Laboratory Analytical Procedures (LAP) established by National Renewable Energy Laboratory[202].

Enzymatic hydrolysis experiments were performed in Prof. Nathan Mosier's Lab (Purdue University, West Lafayette-IN).

7.4 Results

7.4.1 Support Vector Regression

Effect of augmenting the training dataset. The values of the input variables from the 53 lines present in the original database were the mean values reported in the references using anywhere between n = 6 to 18 lines. Ideally, having data on each of the individual lines would have allowed to incorporate natural biological variance into the regression model thereby enabling better predictive performance. In addition, validation performance of a regression model may also be affected by the size of the training data set. Smaller data sets may lead to overfitting for a highly-parameterized model, resulting in poor validation performance. In an effort to account for biological variance and prevent overfitting of the regression model, synthetic data was generated using the standard deviations reported in the references (Table 7.1). Validation performance of SVR models improved with increasing training data size (Figure A4.1) and showed no significant improvement over 750 data points. For all further simulations, a training data set of 725 data points was employed.

Improved performance by oversampling underrepresented data points. The initial data set constituting the 53 lines did not span the range of all the input variables, specifically %H and %S lignin (Figure A4.2). A total of 43 lines corresponded to a %H lignin of lower than 10%, while only 12 lines spanned the range from 20-100%. Similarly,

only one line had a %S lignin higher than 90% and none in the range of 50-80%. Such a skewed distribution of input variables in the training data set may result in a biased model that may lead to poor validation performance[203]. In classification problems, cases of imbalanced data sets have been addressed by (i) oversampling the minority class[204], (ii) undersampling the majority class[205], or (iii) by a modified SVM algorithm where the misclassification penalty (ξ_i) of the underrepresented class is larger[206]. The oversampling strategy was employed to ameliorate the imbalance in training data set and improve the regression performance. Twice the number of data points were generated for the 12 underrepresented lines when expanding the data set. Initial SVR models trained on weighted data sets resulted in a better validation performance over the data sets without weighting (Figure A4.3).

SVR models of bootstrapped training data sets. SVR models for %saccharification efficiency and plant height were obtained for each of the 500 training data sets sampled using empirical bootstrapping (Materials and Methods). The SVR model framework was setup using the Gaussian kernel and was subjected to a *k*-fold (*k*=10) cross-validation procedure. Predictions from all 500 models were plotted against the original target variables to evaluate the regression performance on training data. SVR models for both %saccharification efficiency and plant height were in agreement with the measured data with high correlation coefficients (Figure 7.2 (a)&(b) and Table 7.2). The SVR models were then validated by evaluating their performance on an independent set of *synthetic* data that was not included in training. Predictions for both the target variables were highly correlated with the experimental measurements (Figure 7.3 (a)&(b) and Table 7.2). A significant reduction in validation performance (R~0.71) over training (R~0.96) was

observed for models predicting plant height. Model predictions saturated at plant heights of 45 cm and deviated at lower heights (Figure 7.3 (b)). Paucity of data at lower heights maybe a major cause for a reduction in performance.

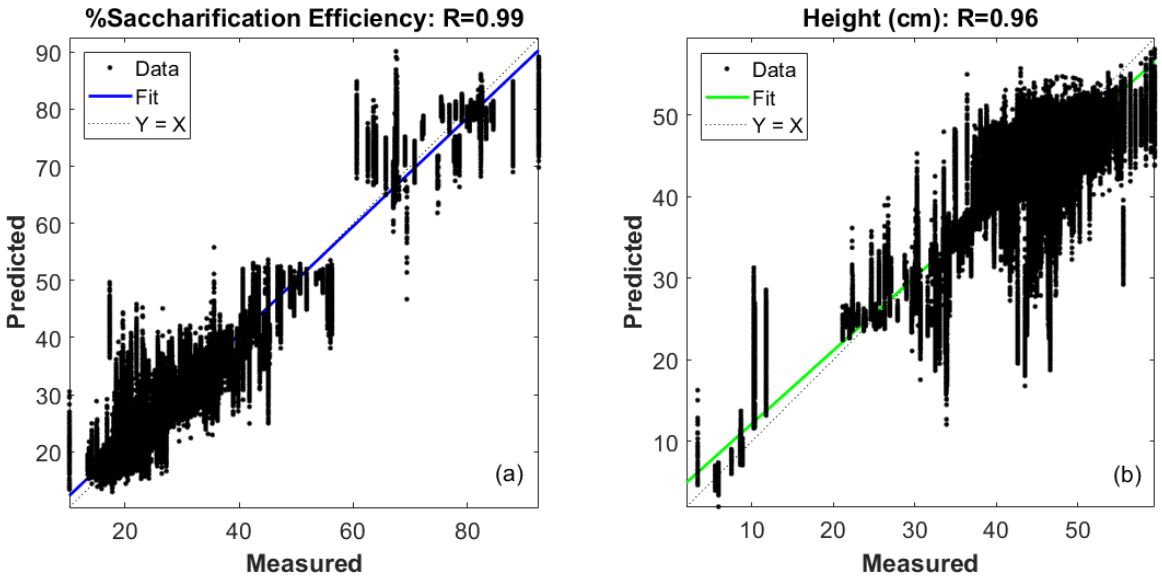


Figure 7.2: Performance of the SVR models on the training data in predicting %Saccharification efficiency (a) and plant height (b). Predictions from 500 SVR models corresponding to the 500 training data sets sampled using empirical bootstrap were combined.

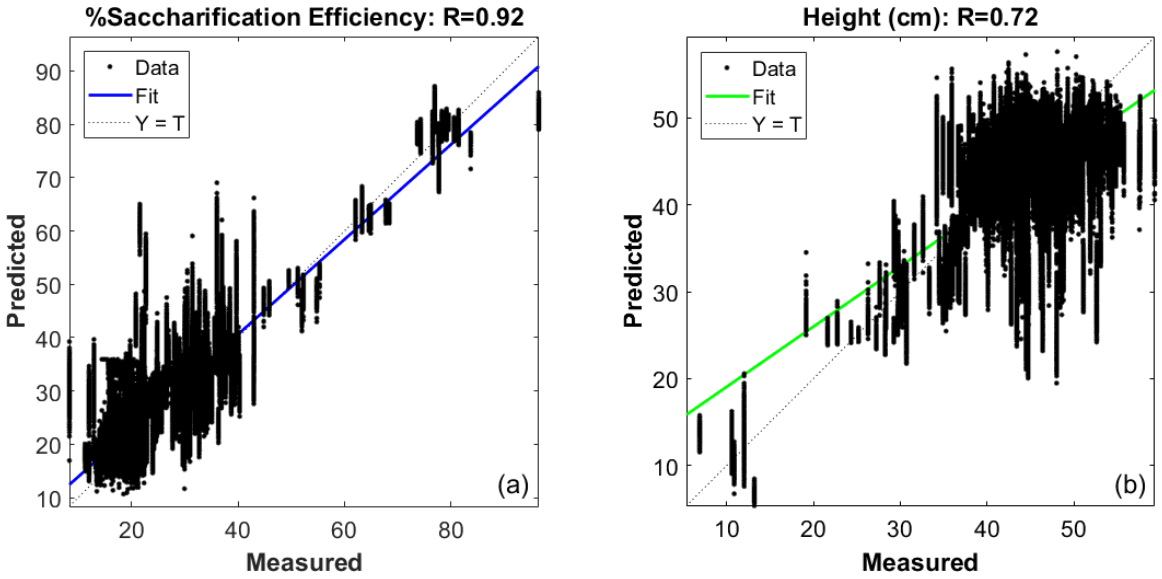


Figure 7.3: Performance of SVR models on the validation data set in predicting %Saccharification efficiency (a) and plant height (b). Predictions from 500 SVR models were combined.

Table 7.2: Performance statistics of SVR models on training and validation data sets.

	%Saccharification efficiency		Plant height (cm)	
	mse[a]	R[b]	mse	R
Training	15.9	0.99	13.4	0.96
Validation	31.3	0.92	38.4	0.72

[a]mean square error
[b]correlation coefficient

Hyperparameter optimization of SVR models resulted in a very poor validation performance. The SVR framework has three hyperparameters – also called meta-parameters – namely box constraint (C), sensitivity window (ε), and kernel scale (γ) that can be optimized to obtain superior regression performance. This was achieved by using Bayesian optimization where the *k*-fold (k=10) cross validation loss of regression was minimized by varying C, ε, γ. Significant improvement in regression performances for both %saccharification efficiency (R~0.99) and plant height (R~0.99) was observed on the training data as a result of this effort (Figure A4.4(a), (b) and Table A4.1). Although the models were well trained, they performed poorly on the validation data set (Figure A4.5 (a), (b) and Table A4.1) compared to the models with default values of the hyperparameters. Such poor performance maybe attributed largely to overfitting on the training data[207] and this was apparent in the prediction of plant height where a majority of models predicted a plant height of around 45 cm across the entire range of the experimental measurement (Figure A4.5(b)). A more uniform distribution of data would have ensured the presence of a significant number of support vectors for smaller plant heights leading to better regression performance. In the interest of retaining generalizability and accurate prediction capabilities, all SVR models considered for simulations hereafter had default

hyperparameter values assigned by the *fitrsvm* function in MATLAB (see Materials and Methods).

7.4.2 Validation of SVR Model Predictions

The validation performance described in the previous section was on a subset of the augmented data set. The data set – though not trained on any SVR models – does not truly represent an independent set as it was generated by randomly sampling from a distribution on the original data set of 53 lines. The true test of validation for the SVR models would be to successfully predict %saccharification efficiency and plant height on lines that were not among the 53 lines in the original data set. To achieve this, two Arabidopsis mutant lines *c4h-2* and *c4h-3* from Van Acker et al., 2016[35] were set aside for validating the SVR models. These two lines represented a unique H:G:S composition with almost no H derived lignin, and a G:S composition close to 1:1 (Table A4.2). Deficient in cinnamate-4-hydroxylase, *c4h-2* line has a dwarfed phenotype and a high saccharification efficiency while the line with a mutation in *c4h-3* grows to wild type height with a saccharification efficiency slightly higher compared to wild type plants. In addition to these lines, the SVR model was validated on *4cl1 4cl3 4cl4* and *4cl1 4cl2 4cl3* triple mutant lines[22]. Both mutant lines are deficient in three different isoforms of 4-coumarate ligase (4CL) catalyzing the conversion of *p*-coumaric acid to *p*-coumaroyl CoA in the phenylpropanoid pathway, a reaction common to all three monolignols, yet the resulting phenotypes are significantly different. Although both mutant lines are characterized by almost similar H:G:S composition, the *4cl1 4cl2 4cl3* triple mutant exhibits a dwarfed phenotype (Figure 7.3Table A4.4). Taken together the validation set consisting of the four lines mentioned

above offered a wide variation in the input and target variables allowing for rigorously testing the trained SVR models.

Table 7.3: Experimental measurement and SVR model prediction on wild-type and transgenic lines.

Line	%Saccharification Efficiency		Height (cm)		Reference
	Measured[a]	Predicted[b]	Measured[a]	Predicted[b]	
c4h-2	50.5 ± 2.4	50.3 ± 0.4	35.4 ± 4.1	31.8 ± 0.5	[35]
c4h-3	24.1 ± 2.1	23.8 ± 1.9	49.9 ± 2.6	48.5 ± 1.6	[35]
4cl14cl34cl4[a]	42.0 ± 1.7	34.9 ± 9.6	38.9 ± 6.1	44.4 ± 2.4	[22]
4cl24cl24cl3[a]	28.3 ± 0.8	25.6 ± 5.7	15.8 ± 2.4	45.0 ± 2.2	[22]

[a] Data presented as mean ± S.D. over n=3 biological replicates
[b] Data presented as mean ± S.D. from the distribution of solutions obtained from running 500 SVR model pairs.
[c] Saccharification efficiencies were scaled as described in Materials and Methods.

The SVR models successfully predicted saccharification efficiency and plant heights for *c4h-2* and *c4h-3* lines. Although both lines had similar H:G:S composition [35], the model was able to predict the dwarfed phenotype due to significantly reduced lignin phenotype exhibited by *c4h-2* lines [35,114]. Interestingly, in case of the triple mutants the SVR model was able to predict both the saccharification and growth phenotype for *4cl1 4cl2 4cl4* lines, but failed to predict the dwarfed phenotype for *4cl1 4cl2 4cl3* line. Both triple mutants have the same total lignin content and %H lignin with a slight different in G and S composition. The failure of the model to predict the growth phenotype for the latter mutant line indicates lack of sensitivity to changes in G and S composition under dwarfed conditions, which in turn stems from a lack of data corresponding to dwarfed lines. Also,

recent studies have shown that dwarfism in lignin-deficient plants may be linked with sensing of reduction or hyperaccumulation of a metabolic intermediate, or due to changes in cell wall characteristics[156], factors that have not been incorporated into the SVR models. Expanding the set input variables would allow for more accurate predictions of saccharification and growth phenotypes.

7.4.3 Optimization of Total Saccharification Yields using Genetic Algorithms

For lignocellulosic feedstock to be transferable to the field, in addition to a high saccharification efficiency the biomass yields ought to be higher. Hitherto genetic engineering experiments have focused on how altered lignin content and composition affects biomass digestibility, but the pleiotropic effects of such manipulations haven't been well understood. Dwarfism, slower growth rates, sterility, collapsed xylem vessels are some phenotypes observed in certain transgenic lines. There is clearly a trade off in reducing the total lignin content in plants given its importance for supporting plant vasculature and upright growth. In this study, the optimum of this trade-off was investigated by formulating an objective function that considered the effect of biological modifications on both saccharification efficiency and plant height. Although plant weight would better represent the biomass yield, plant height was considered for the study due to the large variation in the measured weight reported across different studies.

GA were used to estimate the optimal lignin content and composition in terms of %H, %G and %S lignin that maximizes the total saccharification yield. The objective function (fitness function) corresponding to the total yields was represented as a product of %saccharification efficiency and plant height. This can be envisioned as a product of the

functional forms obtained for %saccharification efficiency and plant height as a result of SVR ($f_1(\mathbf{x}) \times f_2(\mathbf{x})$; see Materials and Methods). Each pair of the 500 pairs of SVR models generated using empirical bootstrapping was optimized using GA.

This effort resulted in two optimal solutions (Table 7.4), each with a moderately reduced lignin content but a different H:G:S composition. Solution one indicated a 51% reduction in lignin compared to wild type lines and almost an even distribution of H, G and S derived lignin. The phenotype was predicted to have a high saccharification efficiency, although slightly dwarfed relative to wild type plants. Interestingly, phenotype corresponding to Solution 1 is similar to that of *cse2* lines of Arabidopsis ([25], Table 7.4) that accumulate higher H lignin units and are reported to have very high saccharification efficiencies. Solution 2 presented a more moderate phenotype with only a 32% reduction in lignin and a higher S/G ratio (0.78). The line was predicted to grow to wild type height with over a 3-fold higher %saccharification efficiency. Although resembling the *c4h-2* and *c4h-3* lines in terms of the relative G and S derived lignin (Table 7.4), the higher total lignin content, absence of H derived units taken together with the high saccharification efficiency and near wild-type growth phenotype makes this solution a unique prediction of the SVR-GA model.

Table 7.4: Optimization results obtained from the SVR-GA methodology

Line/Solution[a]	%AcBr lignin	%H lignin	%G lignin	%S lignin	%Sacc. Eff.	Height (cm)	Fitness function value
1.	9.6 ± 0.9	30 ± 4	39 ± 3.5	31 ± 2	67 ± 5.8	38 ± 4	2545 ± 154
2.	13 ± 0.2	0.1 ± 0.1	56 ± 0.6	44 ± 0.6	58 ± 2.8	44 ± 1.6	2560 ± 121
Wild type [b]	19 ± 4.3	1.5 ± 1.4	66 ± 6.2	32 ± 5.8	18 ± 2.4	46 ± 3	844 ± 123
cse2[c]	11.3 ± 0.4	27 ± 1.1	39 ± 0.7	34 ± 0.7	78 ± 7	25 ± 3	1950 ± 290

[a] Data presented as mean ± S.D. from the distribution of solutions obtained from running 500 SVR model pairs.
[b] Data presented as mean ± S.D. using all wild-type plants included in the original data set.
[c] Data represented as mean ± S.D from biological replicates reported in [25]

7.5 Conclusions

In this study, SVM based regression was used to predict saccharification efficiency and plant height of Arabidopsis as a function of total lignin content and lignin composition. Experimental data, constituting 53 lines from 9 across independent studies, was expanded to achieve improved performance of the regression scheme for more accurate predictions. This effort resulted in SVR models for saccharification efficiency (R~0.92) and plant height (R~0.73) that were in close agreement with the experimental data in the validation set. Confidence intervals for the predictions were obtained using an empirical bootstrap sampling technique where 500 data sets were generated and trained using SVM for both the response variables considered for the study. Optimization by GA of the trained SVR models resulted in two solutions that maximized the net saccharification yield. As expected,

both solutions resulted in phenotypes with reduced total lignin content. Interestingly, the solutions had strikingly different lignin compositions with one having an even distribution of the H, G, and S derived units while the other was made virtually of G and S lignin with an S/G ratio close to 0.78. Future models can be constructed with a more accurate indicator of biomass yields such as weight, biomass density, or stem diameter. Furthermore, incorporating additional biological traits such as %cellulose and hemi-cellulose in cell-walls and matrix polysaccharide composition would improve prediction performance of the growth phenotype.

8. FUTURE WORK

In this chapter, experiments and possible research directions for the future have been summarized based on the results and findings obtained from this work.

8.1 Alternative Route of Caffeic Acid Synthesis

One of the major conclusions from our work on ^{13}C-metabolic flux analysis (Chapter 4) and targeted metabolomics (Chapter 5) studies on Arabidopsis stems is that an alternative route to caffeic acid synthesis exists. Flux estimates from ^{13}C-MFA indicated ~22% of the total input flux going to caffeic acid from *p*-coumaric acid through a route alternative to the traditional one *via* the shikimate esters of hydroxycinnamic acids. Although, this was observed under fed conditions (1 mM of ^{13}C-Phe) where enzymes may see increased concentrations of the substrates, metabolomics data from *4cl1* mutants clearly indicated that phenylpropanoid intermediates can accumulate to high concentrations under physiological conditions (unfed). In addition, a ~3000-fold higher concentration of caffeic acid – a product of the enzyme that has been knocked out – in *cse2* plants only bolsters the hypothesis that there exists an alternative route to caffeic acid synthesis. To further this line of investigation, the following research directions were proposed.

8.1.1 ^{13}C-Metabolic flux analysis of *med5a/5b ref8-1* and *cse2* mutants

Despite the fact that ^{13}C-MFA on WT and *4cl1* plants provide both experimental and modeling evidence of a novel route to caffeic acid, the possibility of *p*-coumaric acid hydroxylation by C3'H in cells that are not primarily lignifying and in which the substrate

accumulates to a high concentration has not been ruled out. The mediator disrupted reduced epidermal fluorescence plants[156], *med5a/5b ref8-1*, provide the ideal background to test the hypothesis. The *med5a/5b ref8-1* lines accumulate higher *p*-coumaryl-shikimate concentrations than *p*-coumaric acid (Chapter 5, Figure 5.1 (a) & (c)) and are characterized by reduced C3'H activity[156,208]. Therefore, with reduced C3'H enzyme activity and higher accumulation of the predominant substrate, it is highly unlikely to see higher isotopic label enrichment in caffeic acid over caffeoyl-shikimate when fed with $^{13}C_6$-Phe, if C3'H is alone hydroxylating *p*-coumaric acid. Arabidopsis plants deficient in CSE (*cse2*, [25]) provide an interesting background for conducting ^{13}C-MFA because *cse2* mutants exhibit significant accumulation of caffeic acid (Figure 5.1), the product of the enzyme that has been knocked out.

8.1.2 Identifying genes that potentially catalyze *p*-coumaric acid hydroxylation in *Arabidopsis thaliana*

Dynamic labeling experiments and MFA on *med5a/5b ref8-1* lines would provide evidence of the presence of an enzyme other than C3'H that is capable of hydroxylating *p*-coumaric acid, or a lack there of. If the former is true, the next step would be to identify the gene that encodes such an enzyme. As a first, we employed *in silico* comparative gene expression analysis to shortlist a set of putative P450 enzymes using the Arabidopsis Information Resource database (TAIR; https://www.Arabidopsis.org). The CoExSearch tool was used to output a list of co-expressed genes in Arabidopsis using genes encoding every monolignol enzyme, including isoforms for PAL, 4CL and CAD, as a query. In addition, the set of all the genes involved in the phenylpropanoid pathway was used as a multi-gene query. The lists were filtered for genes encoding P450 type enzymes resulting in 31

putative candidates amongst which CYP98A3, the gene encoding C3'H enzyme also appeared. The 31 candidates were further reduced by excluding genes whose functions have been previously characterized (e.g. At2g40890 that encodes C3'H). The remaining candidates were arranged in decreasing order of their expression in whole stems and epidermal peels of internodes (basal 0-3 cm stem fragments) and the top 6 candidates have been presented in Table 8.1. The P450 expression system in yeast[209] can be used to investigate the genes listed in Table 8.1 for hydroxylase activity towards *p*-coumaric acid before further biochemical characterization.

Table 8.1: Putative gene candidates obtained from co-expression analysis

S.No	Gene Id	Expression in whole stems (counts)[a]	Expression in stem epidermal peels (counts)[a]	Alias
1.	At3g20100	510	260	CYP705A19
2.	At4g37310	500	350	CYP81H1
3.	At3g26290*	410	150	CYP71B26
4.	At1g13080	400	240	CYP71B2
5.	At4g27710	200	200	CYP709B3
6.	At4g39510	180	40	CYP96A12

[a]Tissue specific expression was obtained using ThaleMine tool on the Arabidopsis Information Portal ([210]).

8.2 Non-aqueous Fractionation of Arabidopsis Stems Fed with Phenylalanine

A recent study has shown vacuolar Phe sequestration in plants as a response to Phe hyperaccumulation. (Lynch et al., 2017; in press). Similar hyper-accumulating conditions

are observed in Arabidopsis stems when fed with high concentrations of $^{13}C_6$-Phe. Although, estimates of sub-cellular distribution of Phe from NAQF studies on stems indicated no significant vacuolar pool, increased concentrations of Phe and other phenylpropanoid intermediates under fed conditions can activate transport into the vacuole (Lynch et al., 2017). In such a scenario, knowledge of the metabolite pool sizes in the vacuole become important in obtaining more accurate flux maps or developing kinetic models of the phenylpropanoid pathway. Application of the NAQF technique to Arabidopsis stems fed with a high concentration of Phe (1 mM) would resolve the relative distribution of Phe and other phenylpropanoid intermediates in different sub-cellular compartments of the cell.

8.3 Identify Gene Deletion or Overexpression Strategies for Phenotypes Predicted by Machine Learning

With genome-scale metabolic models being increasingly available for various organisms, the last decade has seen a spur of computational approaches to predict targets for genetic modifications in an attempt to guide experimental metabolic engineering strategies[211,212]. Optimization frameworks like OptKnock[213] and OptGene[214] predict gene knockouts and knockdowns to optimize the production of a biochemical. In addition to the above mathematical tools, OptReg[215] and EMILiO (Enhancing Metabolism with Iterative Linear Optimization, [216]) are frameworks that allow identification of genes to be up- or down-regulated in strain optimization. Although these tools have been largely applied to microbial systems for overproducing industrially valuable chemicals, they can be translated to higher eukaryotes.

The latest iteration of the genome-scale model for Arabidopsis includes several reactions pertaining to secondary metabolism, specifically reaction involved in phenylpropanoid metabolism and lignin biosynthesis[217]. The computational tools discussed above can be extended to Arabidopsis for identifying gene targets (deletions and overexpressions) that result in the high saccharification phenotype predicted by our work on machine learning (Table 7.4, Chapter 7).

APPENDIX A: SUPPLEMENTARY INFORMATION

A1. Analytical Method Development (I): Supplementary Figures and Table

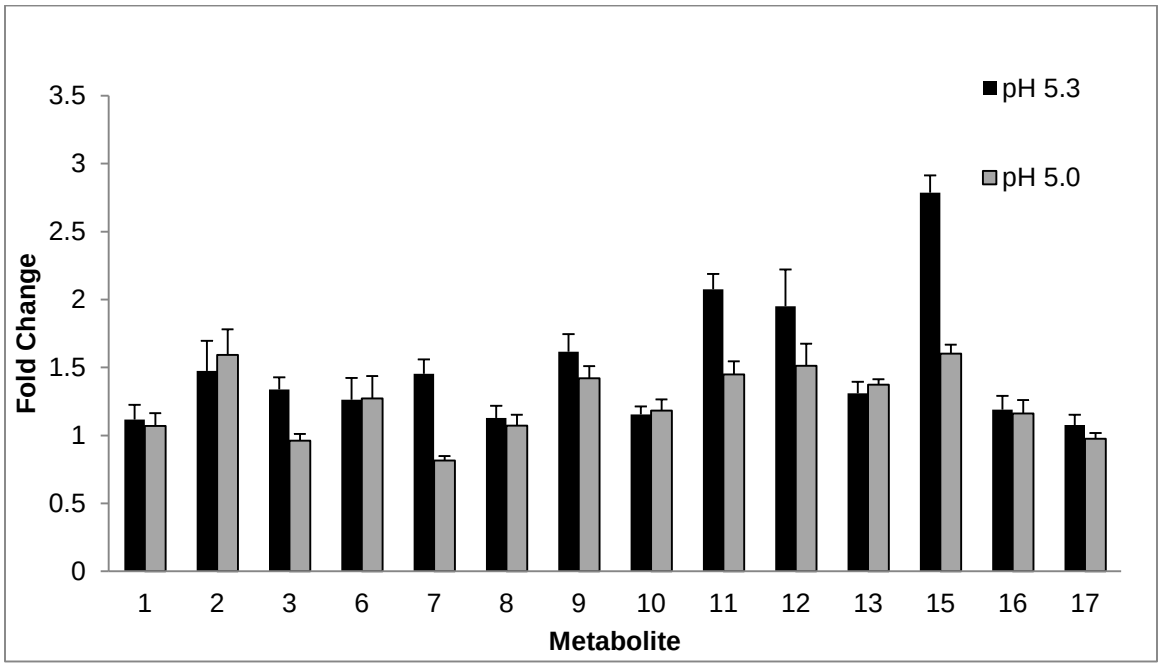


Figure A1.1: Effect of buffer pH on metabolite response. Data presented as fold changes to analyte responses at pH 5.6. Data are means ± s.d. (n=4 replicates). * = $p < 0.05$ and ** = $p < 0.001$ by Tukey's HSD post ANOVA test. Metabolites annotated according to Figure 2.1.

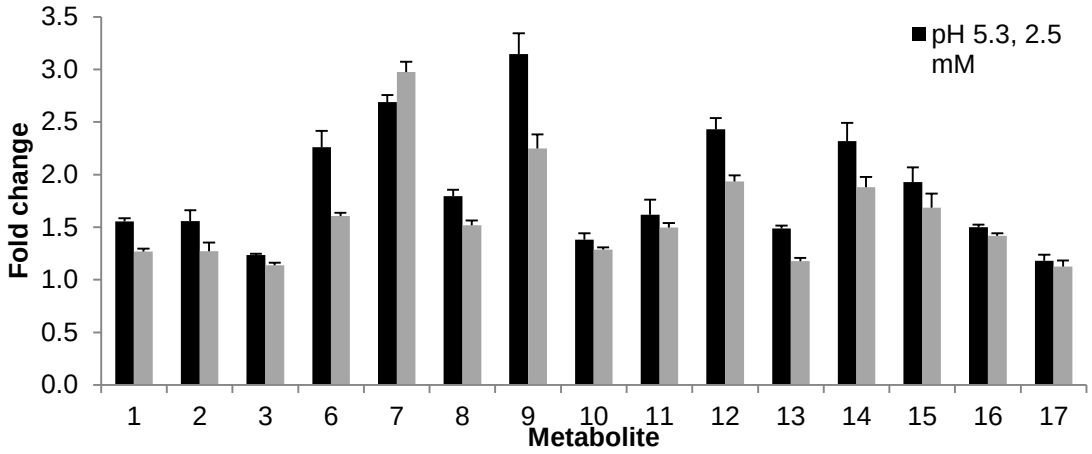


Figure A1.2: Effect of buffer concentration on metabolite response. Data presented as fold changes to analyte responses at pH 5.3, 10 mM. Data are means ± s.d. (n=4 replicates). * = $p < 0.05$ and ** = $p < 0.001$ by Tukey's HSD post ANOVA test. Metabolites annotated according to Figure 2.1.

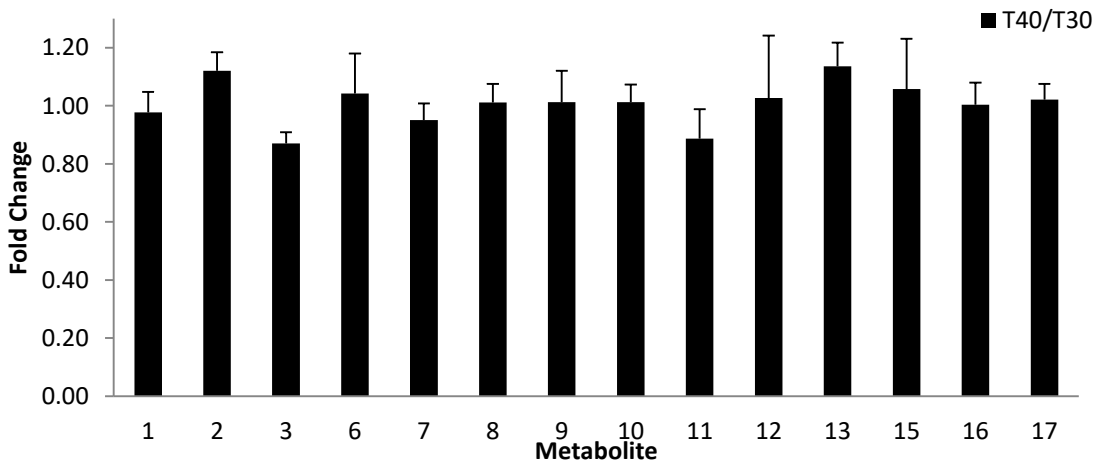


Figure A1.3. Effect of column temperature on metabolite response. Data presented as fold changes of analyte responses relative to column temperature of 30 ^{0}C. Data are means ± s.d. (n=4 replicates). * = $p < 0.05$ and ** = $p < 0.001$ by paired two tailed Student's t-test. Metabolites annotated according to Figure 2.1.

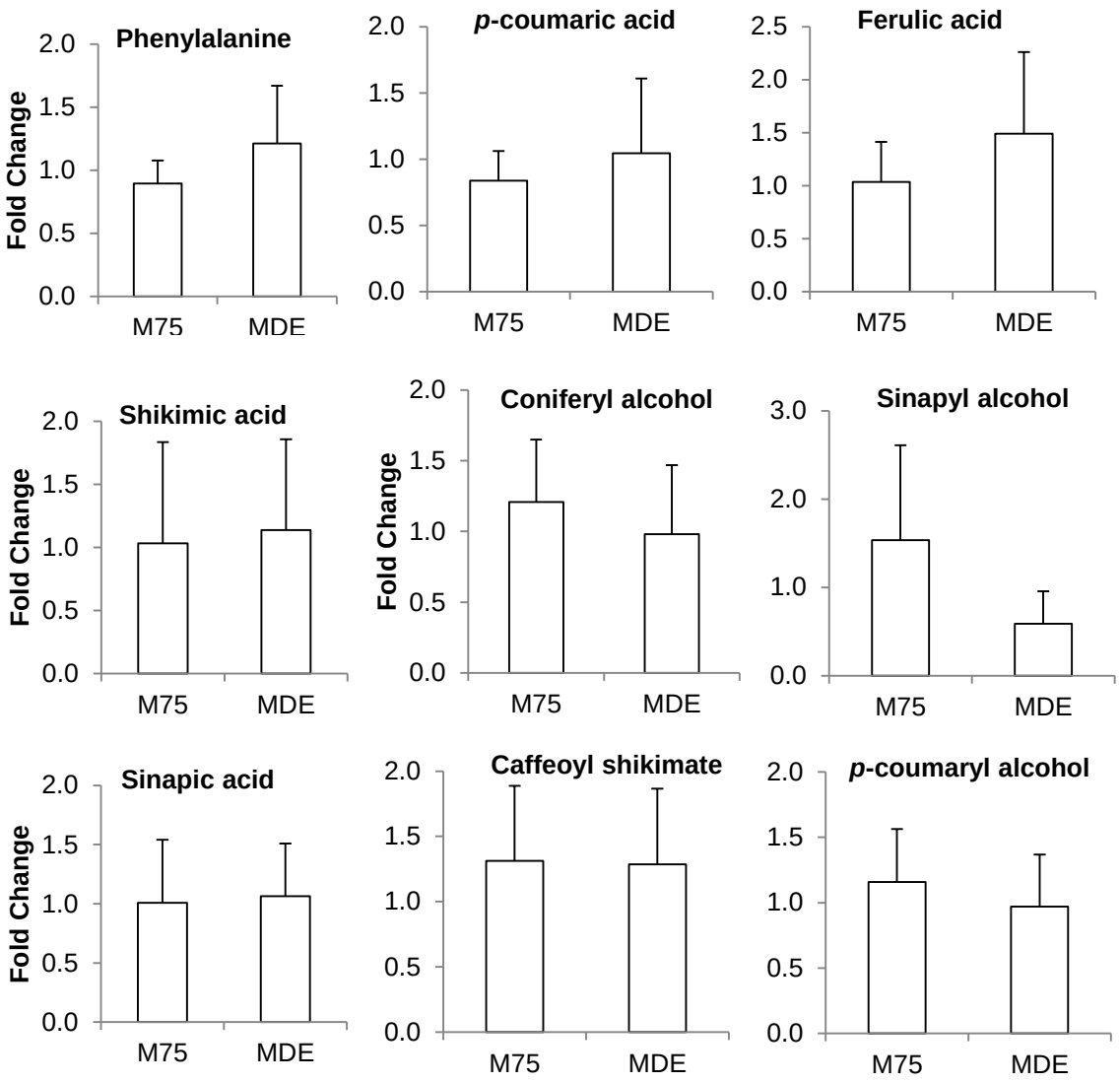


Figure A1.4. Effect of extraction solvent composition on pool sizes. Data presented in fold changes of analyte pools relative to extraction using 50% (v/v) MeOH in water. Data are means ± s.d. (n=4 biological replicates). M75 and MDE denote extraction by vortexing at 25 ^{0}C in 75% (v/v) MeOH in water and double extraction using pure MeOH followed by 50% (v/v) MeOH in water for 60 minutes. ANOVA analysis resulted in no significant differences.

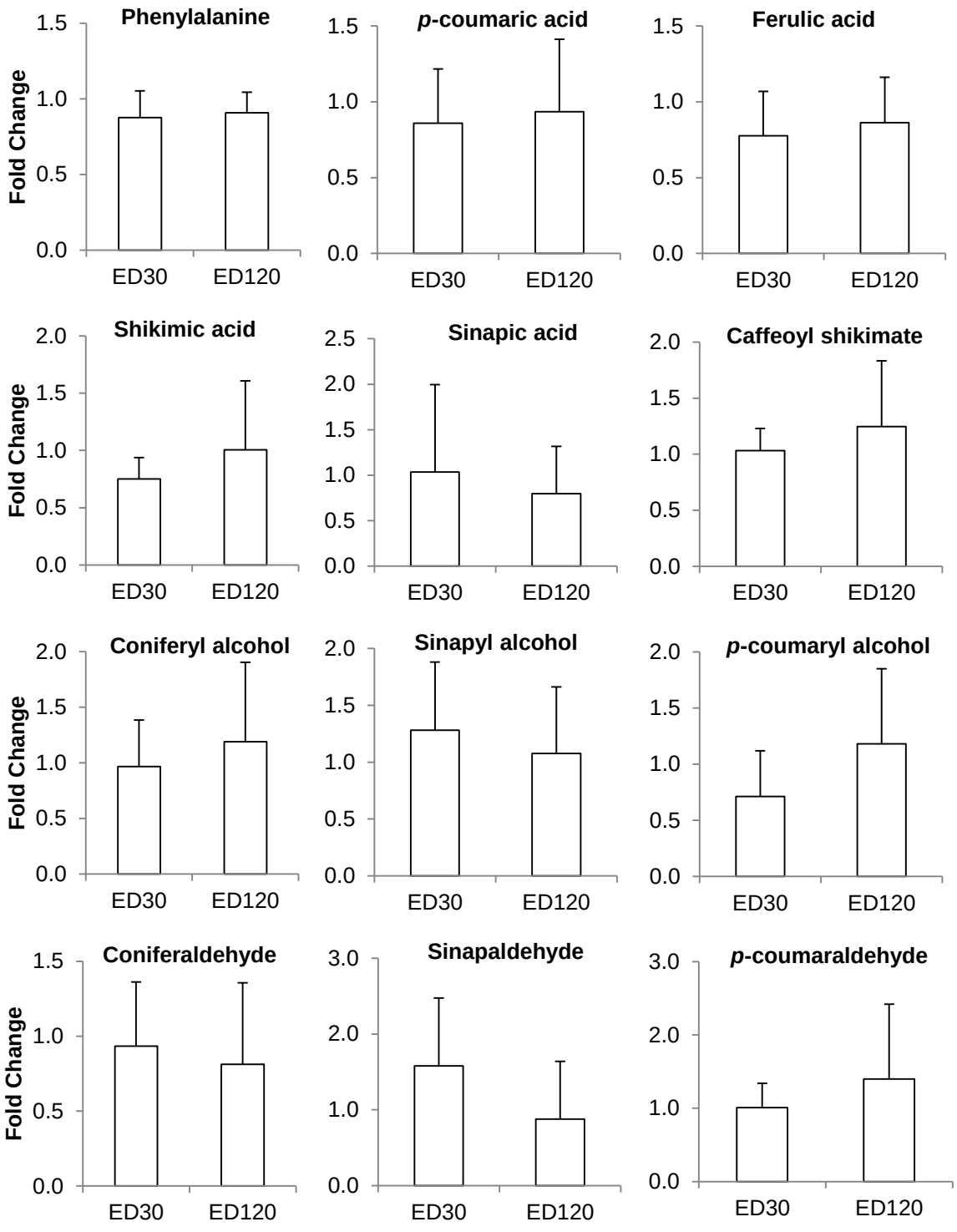


Figure A1.5. Effect of duration of extraction on pool sizes. Data presented in fold changes of analyte pools relative to extraction at 60 minutes. Data are means ± s.d. (n=4 biological replicates). ED30 and ED120 denote extraction by vortexing at 65 ^{0}C in 75% (v/v) MeOH in water for 30 and 120 minutes respectively. ANOVA analysis resulted in no significant differences.

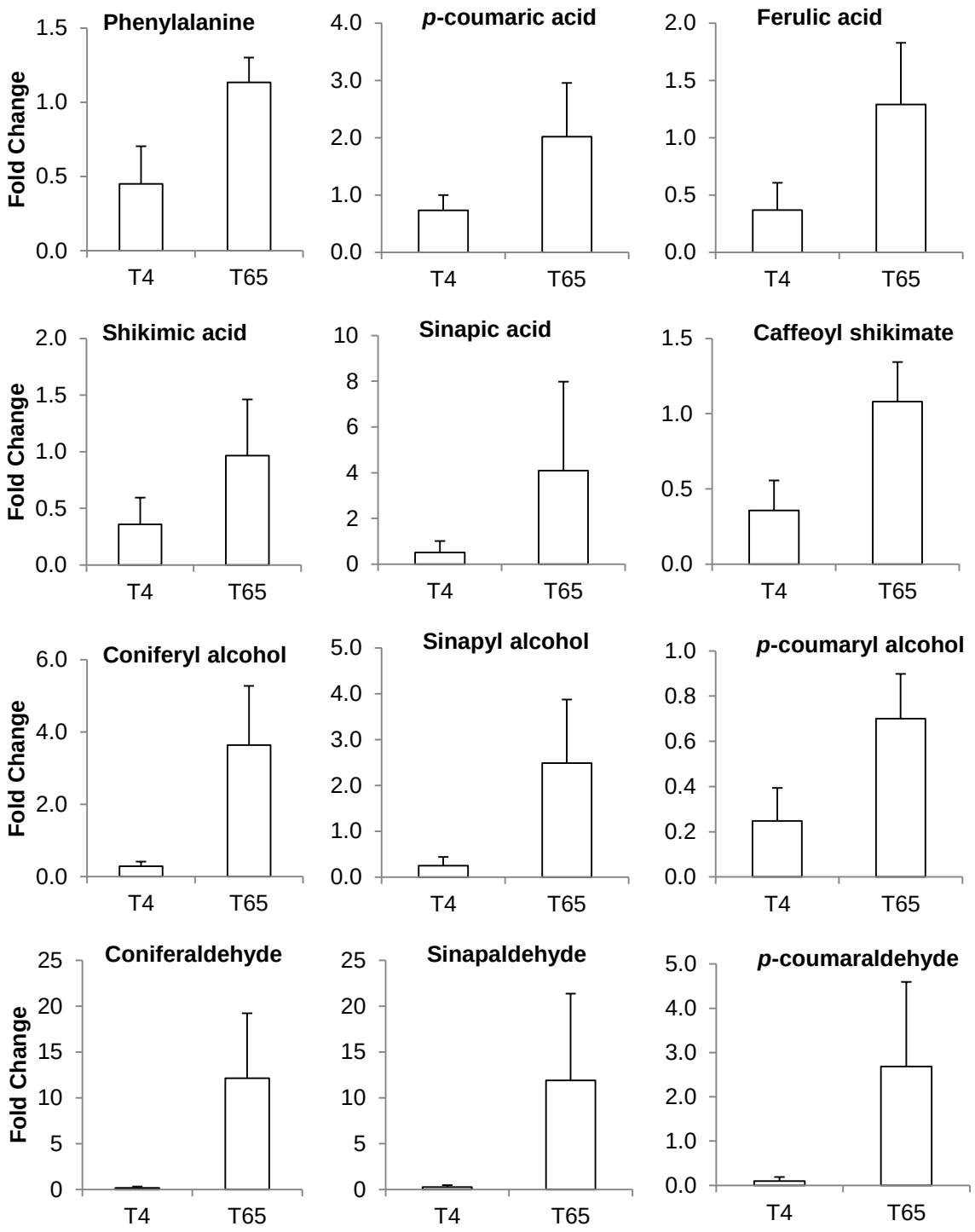


Figure A1.6. Effect of extraction temperature on metabolite pool sizes. Data presented in fold changes of analyte pools relative to extraction at room temperature (T=25 ^{0}C). Data are means ± s.d. (n=4 biological replicates). T4 and T65 denote extraction by vortexing in 75% (v/v) MeOH in water at 4 and 65 ^{0}C respectively.

Table A1.1. Manually tuned QTrap 5500 mass spectrometer parameters for phenylpropanoid pathway intermediates. RT, Retention time; Q1/Q3, Parent Ion Mass/Fragment Ion Mass; DP, declustering potential; EP, entrance potential; CE, collision energy; CXP, cell exit potential.

No.	Metabolite	RT (min)	Q1/Q3	DP (volts)	EP (volts)	CE (volts)	CXP (volts)
1	Phenylalanine	2.53	164.0/147.0	-60	-10	-15	-6
2	Cinnamic acid	16.2	147.0/103.0	-165	-11	-12	-12
3	*p*-coumaric acid	7.13	163.0/119.1	-128	-7	-17	-10
4	*p*-coumaroyl shikimate	7.64	319.2/163.1	-204	-13	-17	-10
5	Caffeoyl shikimate	6.02	335.2/179.1	-212	-11	-18	-8
6	Shikimic acid	1.54	173.0/93.0	-145	-12	-18	-10
7	Caffeic acid	5.16	179.0/135.0	-160	-9	-15	-19
8	Ferulic acid	7.87	193.1/178.1	-100	-8	-15	-8
9	Sinapic acid	8.49	223.1/208.1	-188	-10	-16	-15
10	*p*-coumaraldehyde	11.0	147.0/129.0	-100	-8	-15	-8
11	Caffealdehyde	7.41	163.0/145.0	-260	-11	-23	-15
12	Coniferaldehyde	12.1	177.1/162.0	-168	-9	-16	-11
13	Sinapaldehyde	11.6	207.1/192.1	-125	-10	-19	-21
14	*p*-coumaryl alcohol	6.87	149.1/131.0	-170	-4	-13	-9
15	Caffeoyl alcohol	5.25	165.1/147.0	-205	-7	-16	-12
16	Coniferyl alcohol	7.48	179.1/146.0	-99	-7	-17	-8
17	Sinapyl alcohol	7.23	209.1/194.1	-54	-10	-20	-12
	Feruloyl glucose[a]	5.25	355.2/175.1	-100	-8	-15	-8
	Feruloyl malate[a]	6.57	309.2/193.1	-100	-8	-15	-8
	Sinapoyl glucose[a]	5.52	385.2/205.2	-188	-10	-16	-15
	Sinapoyl malate[a]	7.05	339.2/223.2	-188	-10	-16	-15

[a] Retention times are putative and haven't been confirmed using standards.

Table A1.2. Recovery of metabolites after being subjected to sample preparation protocol using standard mixture at a concentration of 0.01 mg/ml. Data are means ± s.d. (n=4 replicates). * = $p < 0.05$, ** = $p < 0.001$ by applying standard Student's t-test.

		Fraction Recovered	
No.	Metabolite	Average	SD
1	Phenylalanine	0.85	0.26
2	Cinnamic acid	0.88	0.16
3	*p*-coumaric acid	1.06	0.04
4	*p*-coumaroyl shikimate	1.09	0.11
5	Caffeoyl shikimate	1.20	0.09
6	Shikimic acid	1.16	0.08
7	Caffeic acid	1.05	0.05
8	Ferulic acid	1.04	0.03
9	Sinapic acid	1.16	0.09
10	*p*-coumaraldehyde	0.98	0.03
11	Caffealdehyde	1.10	0.08
12	Coniferaldehyde	1.05	0.07
13	Sinapaldehyde	0.98	0.06
14	*p*-coumaryl alcohol	0.86	0.08
15	Caffeoyl alcohol	0.93	0.17
16	Coniferyl alcohol	0.75	0.09**
17	Sinapyl alcohol	0.59	0.05**

Table A1.3. Analysis of Variance (ANOVA) test on the results from the extraction studies. '*p'* value indicates the significance of the effect of the independent variable on the population. Tukey's Honest Significant Difference (HSD) was performed as a post-hoc test to determine the significant differences between groups.

		Extraction Study								
		Extraction Solvent Concentration			Temperature			Extraction Duration		
No.	Metabolite	*p*	M75 vs M50	M75 vs MD	*p*	T25 vs T4	T25 vs T65	*p*	ED60 vs ED30	ED60 vs ED120
1	Phenylalanine	0.28	NS	NS	0.0006	<0.01	NS	0.38	NS	NS
3	*p*-coumaric acid	0.68	NS	NS	0.0515	NS	NS	0.80	NS	NS
4	*p*-coumaroyl shikimate	0.66	NS	NS	0.1911	NS	NS	0.75	NS	NS
5	Caffeoyl shikimate	0.36	NS	NS	0.0005	<0.01	NS	0.54	NS	NS
6	Shikimic acid	0.91	NS	NS	0.0294	<0.01	NS	0.52	NS	NS
8	Ferulic acid	0.23	NS	NS	0.1285	NS	NS	0.40	NS	NS
9	Sinapic acid	0.12	NS	NS	0.0089	NS	<0.05	0.42	NS	NS
10	*p*-coumaraldehyde	0.10	NS	NS	0.0039	NS	<0.05	0.19	NS	NS
12	Coniferaldehyde	0.12	NS	NS	0.0002	NS	<0.01	0.78	NS	NS
13	Sinapaldehyde	0.16	NS	NS	0.0135	NS	<0.05	0.31	NS	NS
14	*p*-coumaryl alcohol		NS	NS					NS	NS
16	Coniferyl alcohol	0.57	NS	NS	0.0007	NS	<0.01	0.72	NS	NS
17	Sinapyl alcohol	0.07	NS	NS	0.0017	NS	<0.05	0.56	NS	NS

Table A1.4. Pool sizes of phenylpropanoid pathway intermediates in WT and *ccr1 A. thaliana* stem tissue. Data of metabolites presented as means ± s.d. (n=4 replicates). Analyte responses normalized to fresh weight of tissue. * indicates $p<0.05$, ** indicates $p<0.001$, *** indicates $p<0.0001$ obtained using the standard Student's t-test. P-value established after the Bonferroni correction is 0.003 indicating that metabolites marked as ** and *** are significantly different.

No[a].	Metabolite	WT	*ccr1*
		Content (nmol/g-FW)	Content (nmol/g-FW)
1	Phenylalanine	25.5 ± 4.73	14.1 ± 2.02*
3	*p*-coumaric acid	0.29 ± 0.06	10.1 ± 1.20***
4	*p*-coumaroyl shikimate	0.07 ± 0.03	0.17 ± 0.03*
5	Caffeoyl shikimate	0.05 ± 0.01	0.09 ± 0.01**
6	Shikimic acid	2.97 ± 0.38	2.59 ± 1.21
7	Caffeic acid	0.05 ± 0.02	0.60 ± 0.07***
8	Ferulic acid	0.14 ± 0.02	32.5 ± 8.17***
9	Sinapic acid	12.0 ± 3.65	24.5 ± 11.9
10	*p*-coumaraldehyde	0.002 ± 0.001	-ND-
12	Coniferaldehyde	0.30 ± 0.04	0.16 ± 0.05*
13	Sinapaldehyde	0.06 ± 0.02	-ND-
14	*p*-coumaryl alcohol	18.3 ± 6.13	-ND-
15	Caffeoyl alcohol	2.53 ± 0.83	4.50 ± 1.62
16	Coniferyl alcohol	3.12 ± 0.37	0.94 ± 0.18***
17	Sinapyl alcohol	23.8 ± 7.82	4.86 ± 2.38**

[a]Metabolite annotation as represented in Figure 2.1.

A2. Metabolic Flux Analysis of the Phenylpropanoid Pathway

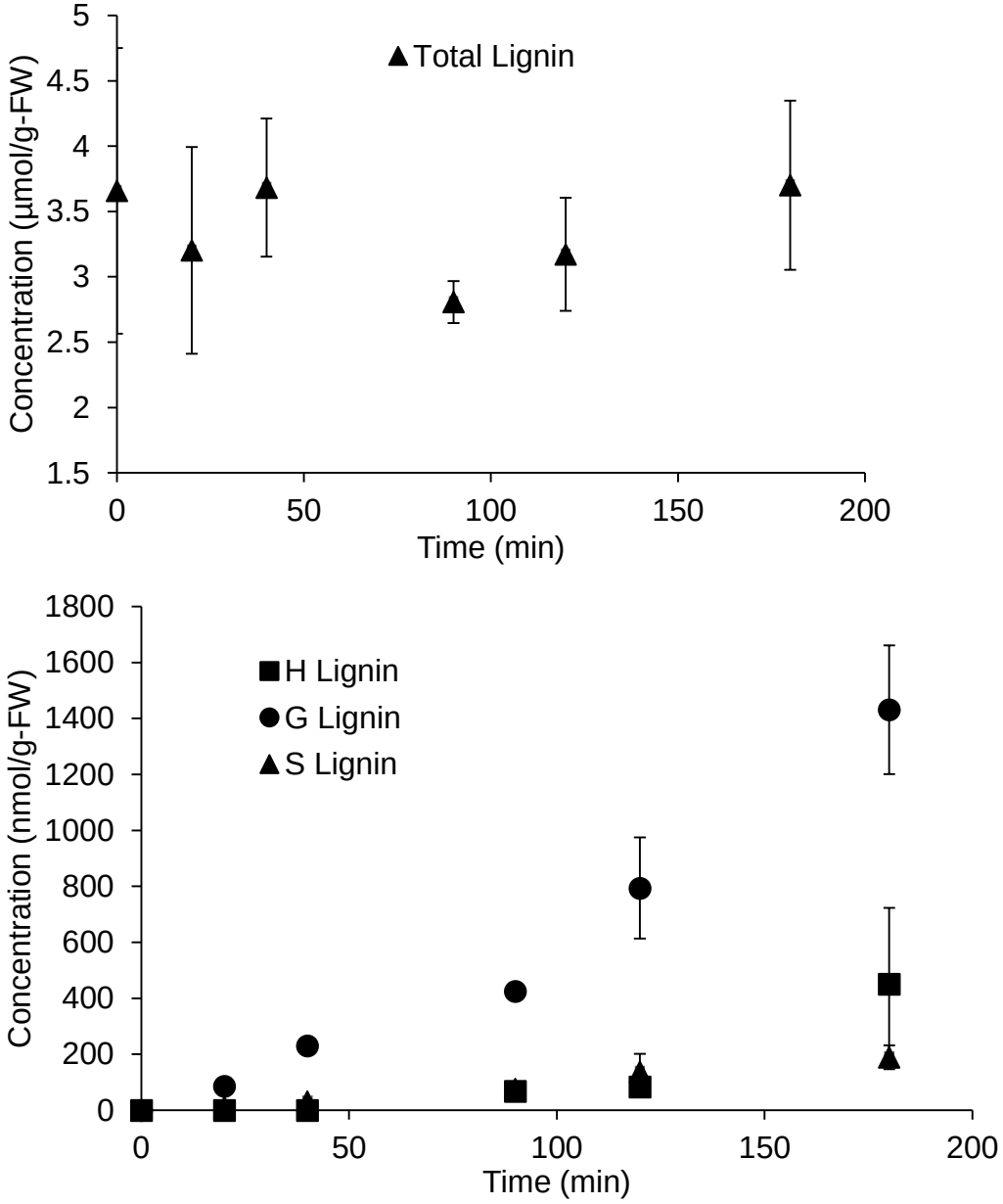


Figure A2.1. Total lignin content (a) and labeled lignin deposition (b) in WT. Marker represent means, and error bars represent standard deviations from *n=3* replicates. Total lignin content was measured using the AcBr method and labeled sub-units of lignin were measured by the DFRC method.

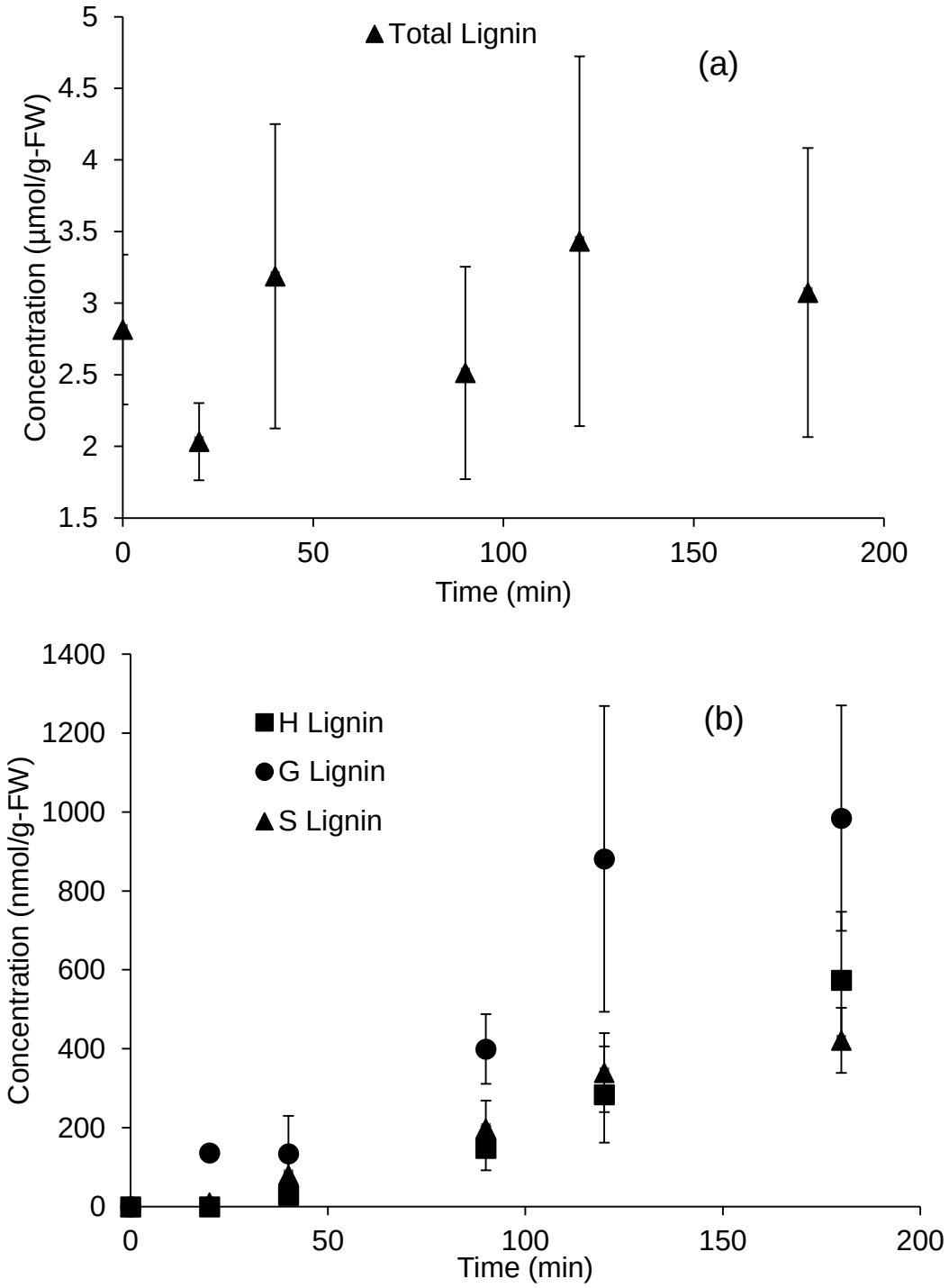


Figure A2.2. Total lignin content (a) and labeled lignin deposition (b) in *4cl1* lines. Markers represent means, and error bars represent standard deviations from *n=3* replicates. Total lignin content was measured using the AcBr method and labeled sub-units of lignin were measured by the DFRC method.

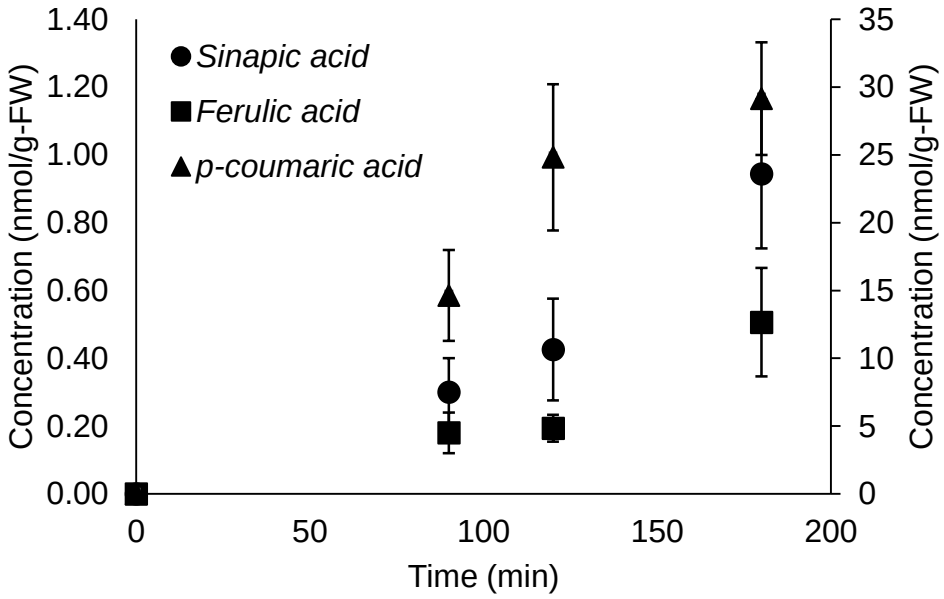


Figure A2.3. Hydroxycinnamic acid accumulation after hydrolysis of extracts from *4cl1* lines. Markers indicate difference in metabolite concentrations before and after acid hydrolysis of extracts. Error bars indicate standard deviations from *n=3* replicates. *p*-coumaric acid has been plotted against the secondary axis (right)

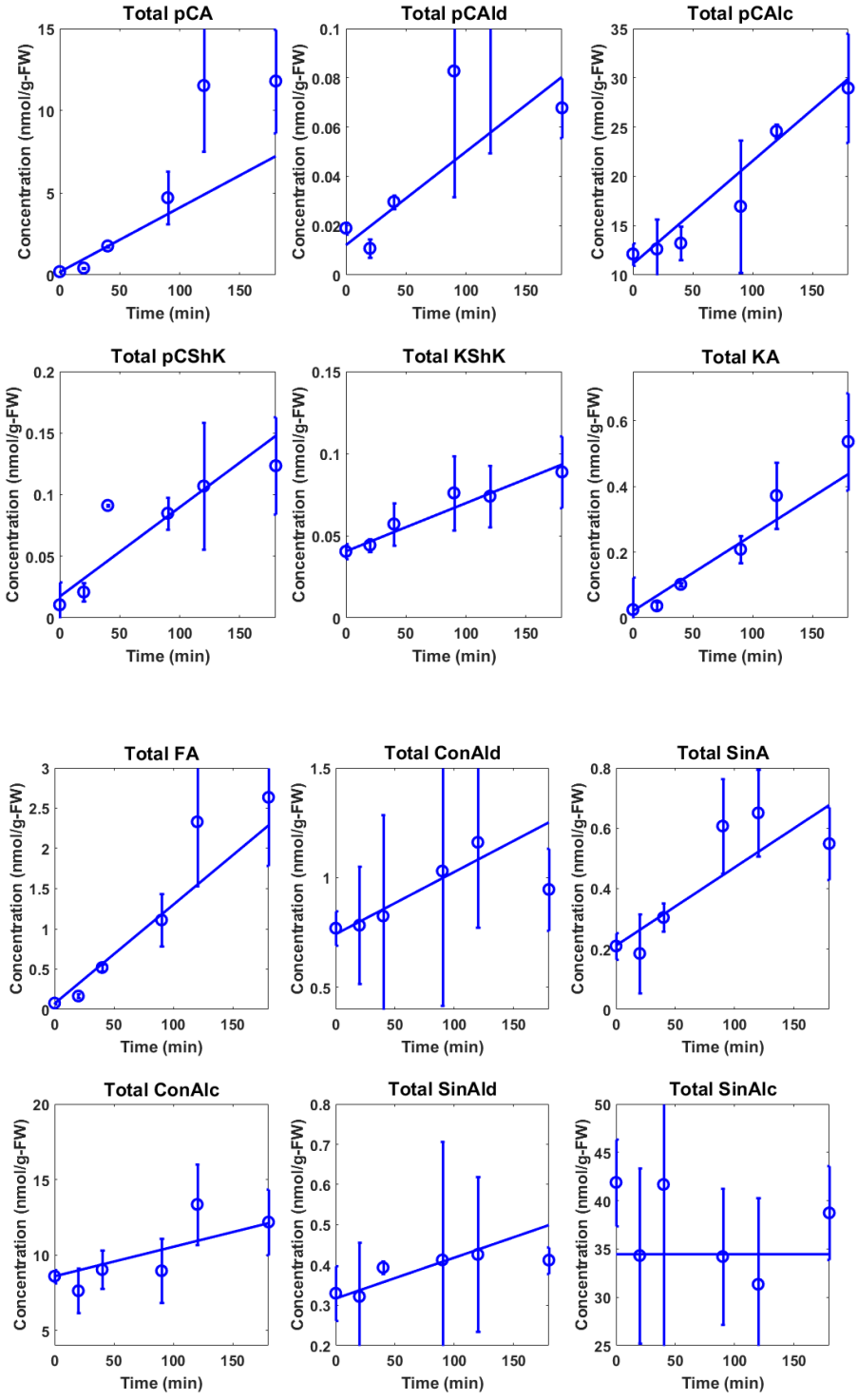


Figure A2.4. Simulated and experimental data of total concentrations and labeled concentrations of metabolites from WT. Markers indicate mean values and error bars are standard deviations obtained from n=3 replicates.

Figure A2.4: Continued

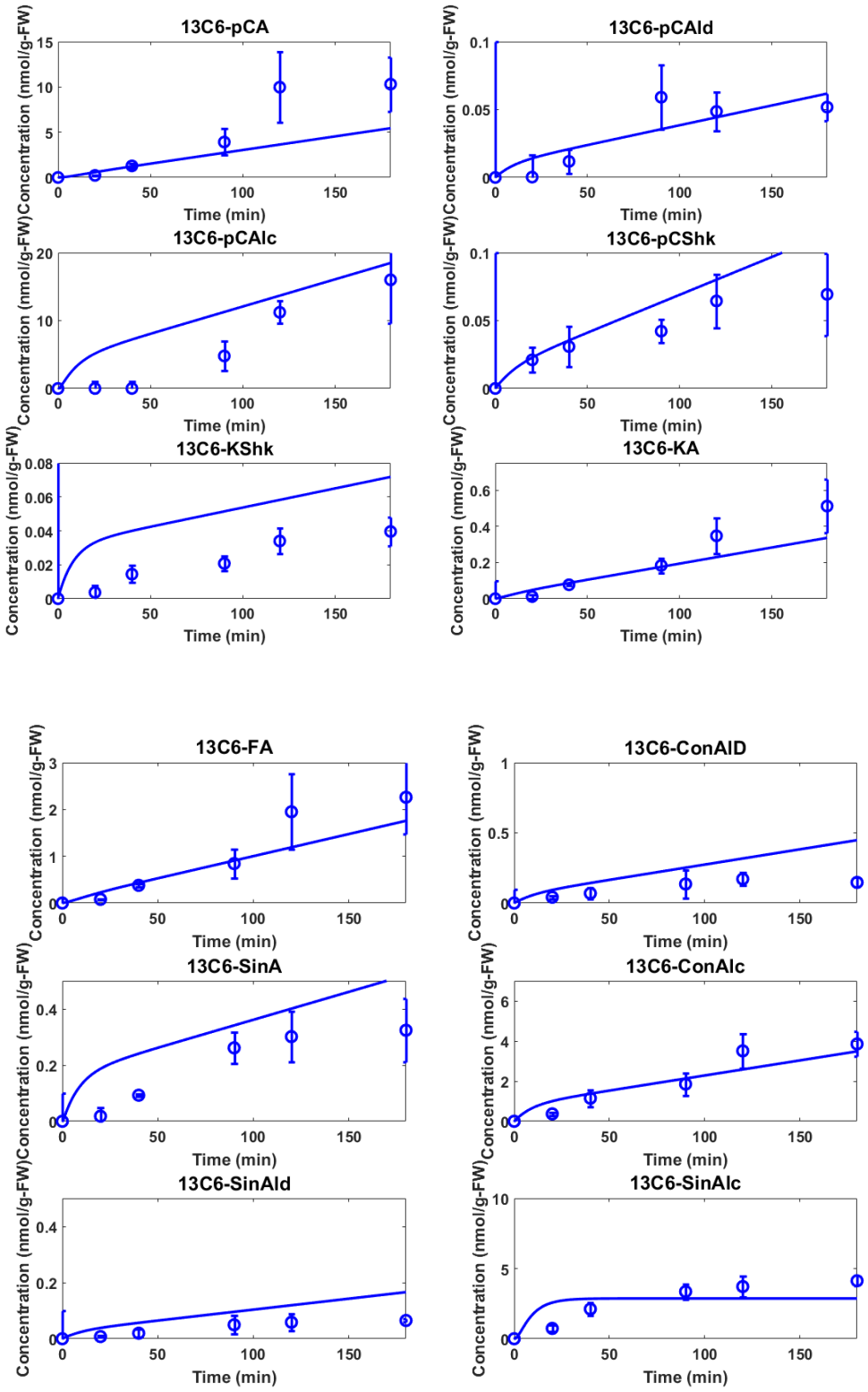
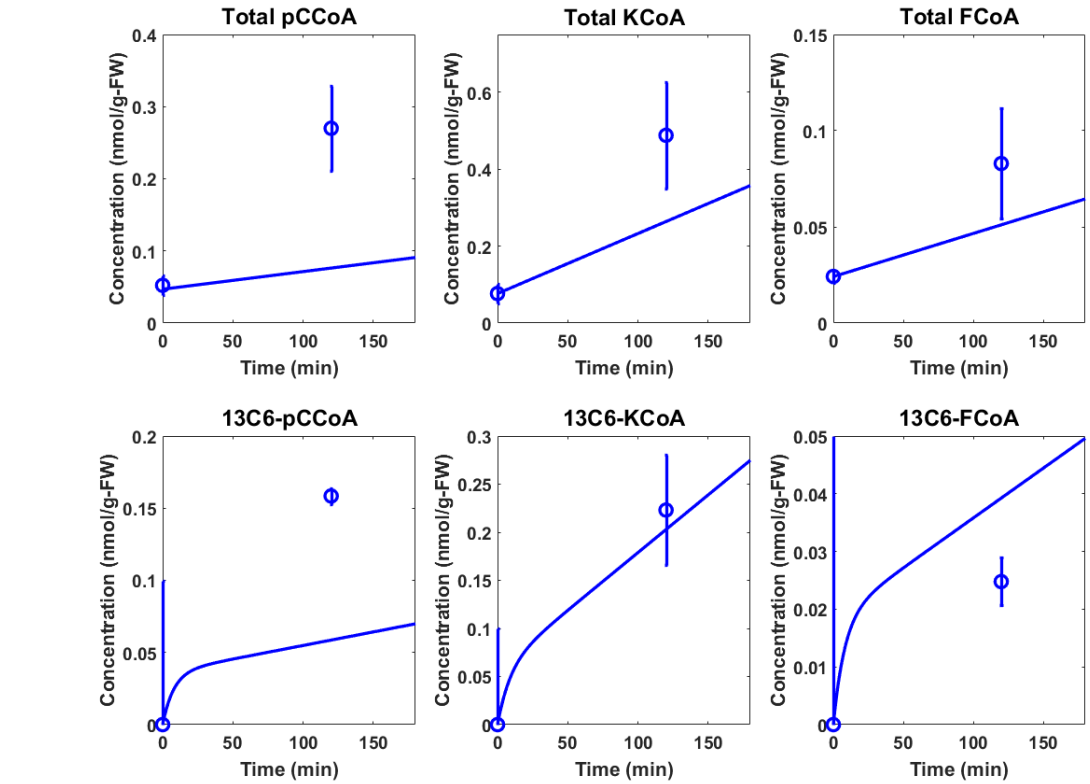
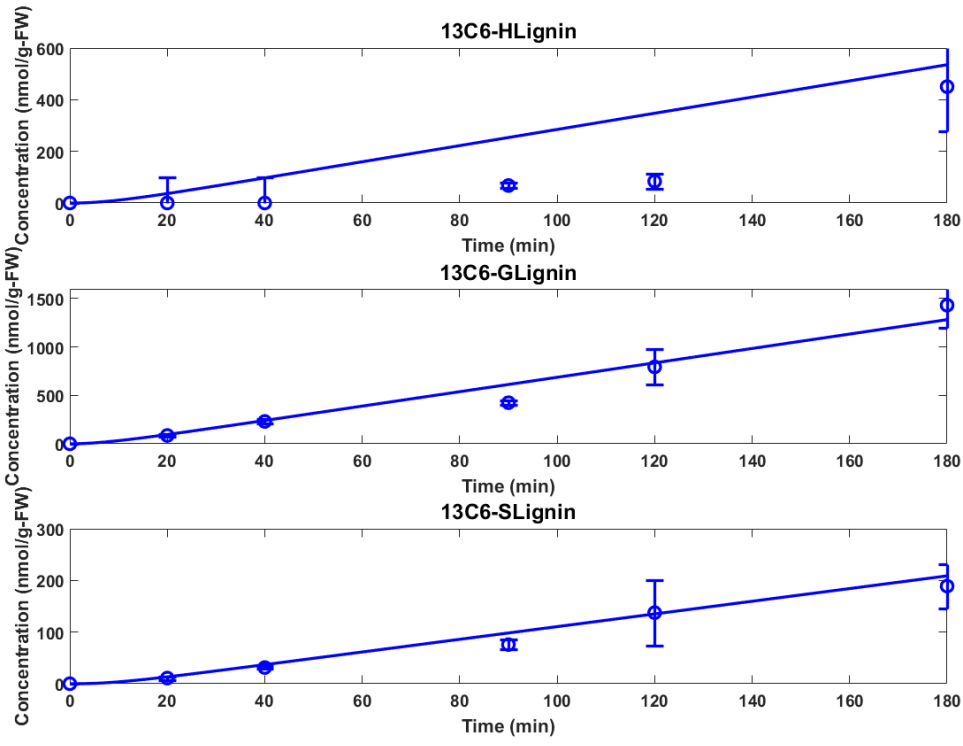


Figure A2.4: Continued





Supplementary Figure 5. Simulated and experimental data of total concentrations and labeled concentrations of metabolites from *4cl1* lines. Markers indicate mean values and error bars are standard deviations obtained from n=3 replicates.

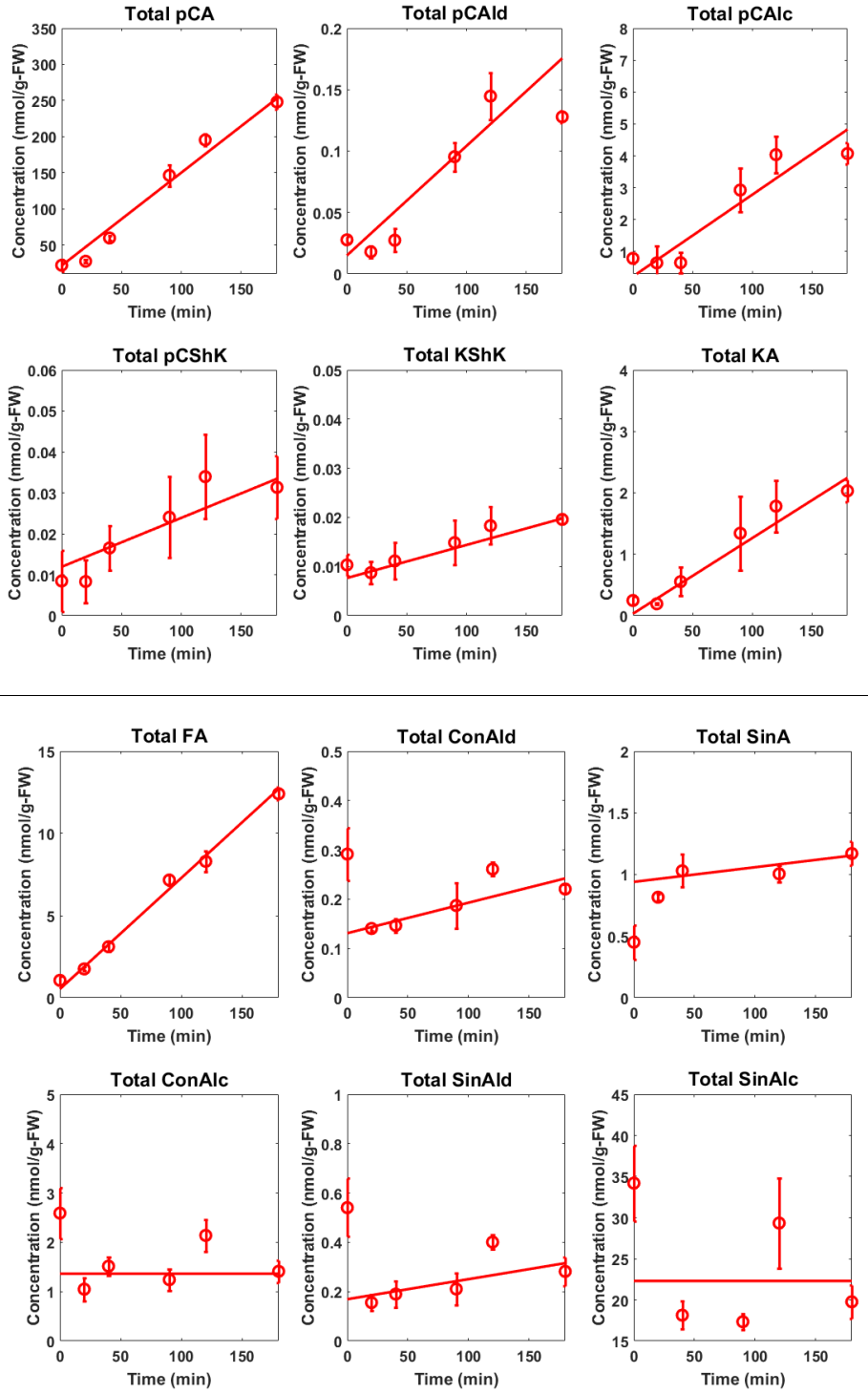


Figure A2.4: Continued

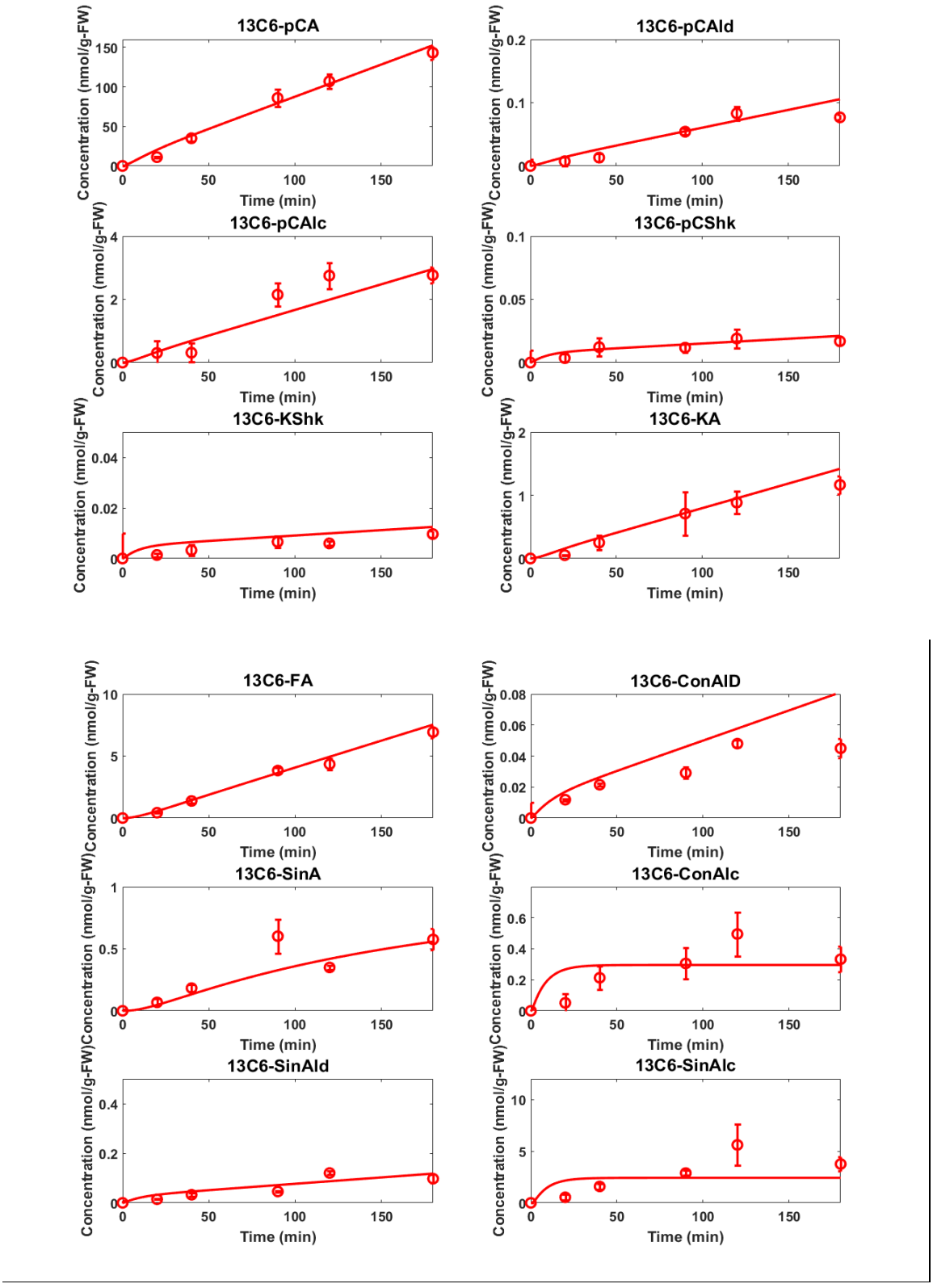


Figure A2.4: Continued

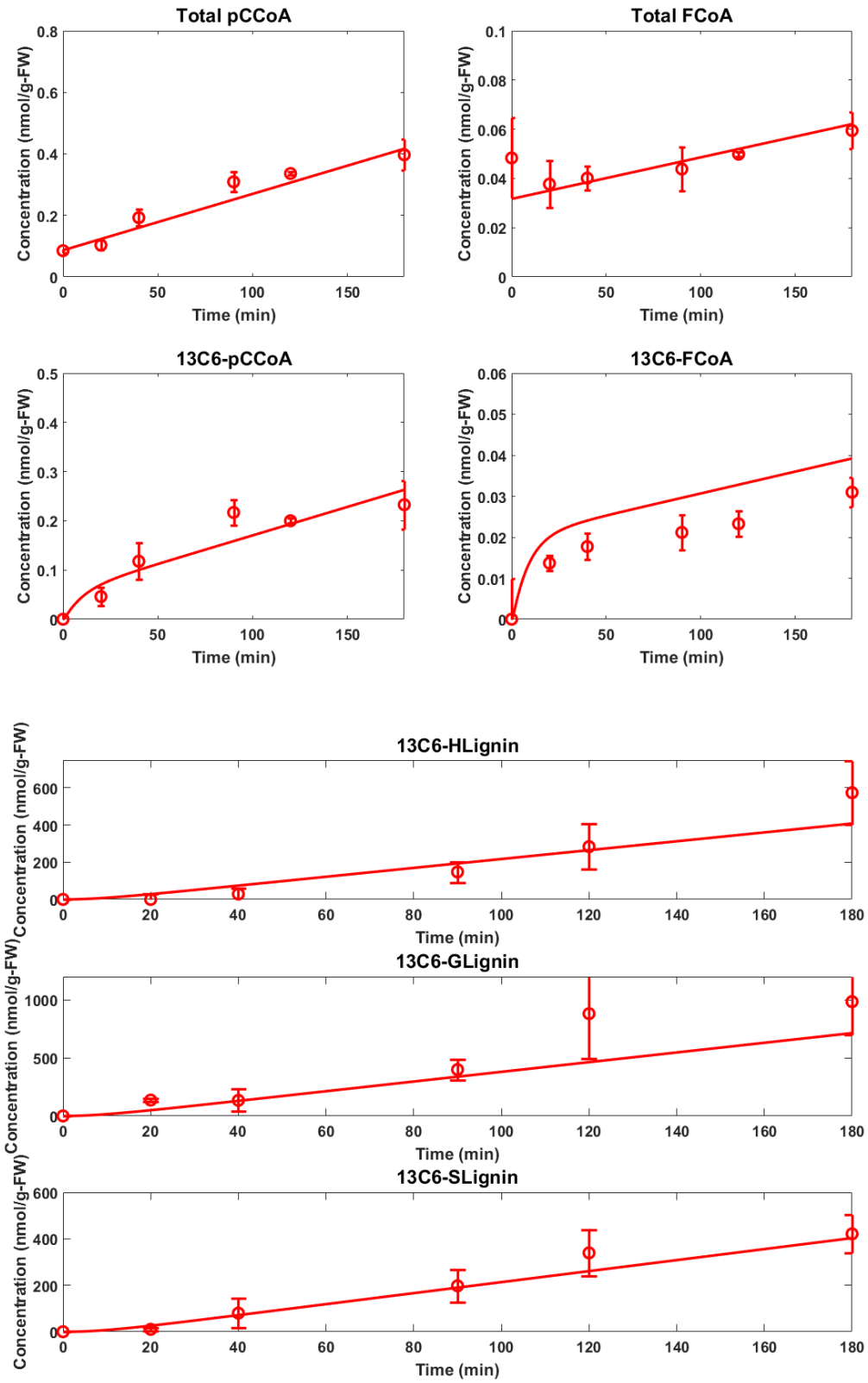


Table A2.1. Mobile phase gradient for analyzing hydroxycinnamyl CoA esters.

Time (min)	Solvent A[a](%)	Solvent B[b](%)
1	90	10
7	10	90
10	10	90
11	90	10
15	90	10

[a]Solvent A: 5 mM ammonium acetate solution buffered to a pH of 6.2 using glacial acetic acid.
[b]Solvent B: 98/2/0.2 (%v/v) of acetonitrile/Millipore water/formic acid.

Table A2.2. Retention time (RT), ion transitions Q1/Q3 (m/z), and ESI parameters for the phenylpropanoid pathway intermediates[a]

Metabolite	RT (min)	Q1 $[M-H]^-$	Q3 $[M-H]^-$	DP (volts)	EP (volts)	CE (volts)	CXP (volts)
p-coumaroyl CoA	4.26	912.3	408.1	-260	-8	-48	-15
Caffeoyl CoA	4.04	928.3	408.1	-260	-8	-50	-17
Feruloyl CoA	4.36	942.3	408.1	-260	-8	-50	-15
Benzoyl CoA	4.15	870.3	408.1	-260	-8	-50	-12

[a] Analysis was performed using an AbSciex QTrap 5500 mass spectrometer coupled to Shimadzu RP-HPLC system.

Table A2.3. Rates of accumulation (slope) and initial metabolite concentrations (intercept) of intermediates from WT and *4cl1* lines. Data reported as means ± S.D of best fit parameters from linear regression.

Metabolite	WT		*4cl1*	
	Mean	S.D.	Mean	S.D.
p-coumaric acid	0.18	0.06	21.45	2.07
p-coumaroyl CoA	0.05	0.01	0.09	0.02
p-coumaraldehyde	0.01	0.00	0.02	0.00
p-coumaryl alcohol	11.12	1.06	0.22	0.07
p-coumaryl-shikimate	0.02	0.01	0.01	0.00
Caffeoyl-shikimate	0.04	0.00	0.01	0.00
Caffeic acid	0.02	0.01	0.03	0.01
Caffeoyl CoA	0.08	0.02	0.00	0.00
Ferulic acid	0.07	0.02	0.54	0.16
Feruloyl CoA	0.02	0.01	0.03	0.00
Coniferaldehyde	0.74	0.04	0.13	0.01
Sinapic acid	0.21	0.03	0.94	0.06
Coniferyl alcohol	8.59	0.11	1.36	0.39
Sinapaldehyde	0.32	0.01	0.17	0.06
Sinapyl alcohol	34.46	2.22	22.30	5.01

Table A2.4. Dynamic label enrichment data in WT. Mean and standard deviations expressed as percentages from *n=3* replicates.

Metabolite	Time (min)					
	0	20	40	90	120	180
Phenylalanine	0.0 ± 0.0	57.6 ± 1.0	59.7 ± 1.1	62.3 ± 4.6	61.7 ± 5.9	63.5 ± 1.5
p-coumaric acid	0.0 ± 0.0	17.1 ± 15.9	60.3 ± 1.9	78.2 ± 3.8	86.8 ± 3.4	90.7 ± 2.3
Caffeic acid	0.0 ± 0.0	26.9 ± 24.5	75.2 ± 1.5	87.7 ± 2.4	92.9 ± 2.0	95.1 ± 1.2
Ferulic acid	0.0 ± 0.0	47.9 ± 6.3	72.6 ± 0.2	74.7 ± 5.9	82.2 ± 6.7	85.4 ± 1.9
Coniferaldehyde	0.0 ± 0.0	5.6 ± 1.2	8.3 ± 0.1	11.8 ± 3.6	15.1 ± 2.0	19.1 ± 0.3
Sinapaldehyde	0.0 ± 0.0	2.7 ± 1.0	5.0 ± 0.2	12.1 ± 1.4	13.7 ± 1.5	15.7 ± 0.8
Coniferyl alcohol	0.0 ± 0.0	4.8 ± 0.1	12.5 ± 1.0	20.4 ± 1.8	26.1 ± 1.7	31.8 ± 3.4
Sinapyl alcohol	0.0 ± 0.0	2.1 ± 0.4	5.3 ± 1.7	10.0 ± 2.0	12.1 ± 2.3	14.6 ± 2.3
Sinapic acid	0.0 ± 0.0	33.3 ± 57.7	22.3 ± 4.8	45.1 ± 14.8	46.7 ± 10.5	58.0 ± 7.0
p-coumaraldehyde	0.0 ± 0.0	3.0 ± 5.2	39.8 ± 7.1	54.6 ± 1.3	63.5 ± 4.8	76.3 ± 1.2
Caffeoyl-shikimate	0.0 ± 0.0	8.4 ± 8.4	27.1 ± 1.8	27.7 ± 2.4	46.2 ± 1.9	45.1 ± 4.8
p-coumaryl-shikimate	0.0 ± 0.0	0.0 ± 0.0	33.7 ± 15.0	49.6 ± 5.0	68.6 ± 28.4	54.7 ± 7.5
p-coumaryl alcohol	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	22.9 ± 20.0	45.7 ± 6.9	53.8 ± 12.9

Table A2.5. Dynamic label enrichment data in *4cl1* lines. Mean and standard deviations expressed as percentages from *n=3* replicates.

Metabolite	Time (min)					
	0	20	40	90	120	180
Phenylalanine	0.0 ± 0.0	77.5 ± 1.8	83.3 ± 1.4	80.9 ± 1.5	81.9 ± 2.1	85.7 ± 0.9
p-coumaric acid	0.0 ± 0.0	40.6 ± 1.5	58.5 ± 3.6	58.7 ± 1.5	54.7 ± 2.0	57.7 ± 1.2
Caffeic acid	0.0 ± 0.0	27.4 ± 3.5	45.1 ± 2.7	52.5 ± 2.2	49.9 ± 3.4	57.2 ± 2.9
Ferulic acid	0.0 ± 0.0	24.6 ± 2.6	44.2 ± 4.0	53.4 ± 0.9	52.2 ± 1.6	55.8 ± 1.9
Coniferaldehyde	0.0 ± 0.0	8.4 ± 0.3	14.7 ± 1.0	16.4 ± 5.2	18.4 ± 0.6	20.4 ± 3.0
Sinapaldehyde	0.0 ± 0.0	9.5 ± 0.9	17.7 ± 2.4	23.3 ± 7.5	30.1 ± 1.6	34.7 ± 1.1
Coniferyl alcohol	0.0 ± 0.0	5.2 ± 5.7	13.8 ± 3.4	24.2 ± 3.9	23.1 ± 4.5	23.5 ± 4.6
Sinapyl alcohol	0.0 ± 0.0	4.0 ± 1.9	8.8 ± 1.1	16.7 ± 1.6	18.7 ± 4.1	18.9 ± 1.8
Sinapic acid	0.0 ± 0.0	8.3 ± 2.9	17.9 ± 3.4	26.5 ± 9.9	34.9 ± 3.8	49.0 ± 3.2
p-coumaraldehyde	0.0 ± 0.0	39.3 ± 31.6	46.5 ± 10.4	57.0 ± 4.5	57.5 ± 2.4	60.1 ± 2.8
Caffeoyl-shikimate	0.0 ± 0.0	15.0 ± 5.0	27.8 ± 8.3	45.5 ± 12.1	32.9 ± 2.0	49.2 ± 4.5
p-coumaryl-shikimate	0.0 ± 0.0	32.0 ± 28.4	70.0 ± 18.8	51.3 ± 9.5	56.5 ± 13.5	54.7 ± 8.7
p-coumaryl alcohol	0.0 ± 0.0	37.7 ± 32.7	42.6 ± 36.8	74.2 ± 6.5	68.1 ± 0.9	67.9 ± 0.4
p-coumaryl CoA	0.0 ± 0.0	43.3 ± 11.5	60.8 ± 14.7	70.2 ± 1.0	59.5 ± 0.6	58.3 ± 7.1
Feruloyl CoA	0.0 ± 0.0	37.3 ± 6.3	44.9 ± 13.4	48.4 ± 2.4	46.7 ± 6.0	52.1 ± 1.3

Table A2.6. List of metabolites for which inactive pools have been invoked when estimating fluxes in WT and *4cl1* lines.

WT		*4cl1*	
Metabolite	Reason	Metabolite	Reason
Phenylalanine	Plastidial pool	Phenylalanine	Plastidial pool
p-coumaric acid	L.E than caffeic acid	*p*-coumaric acid	L.E than *p*-coumaroyl CoA
Coniferaldehyde	L.E than coniferyl alcohol	Coniferaldehyde	L.E than sinapaldehyde
p-coumaraldehyde	-	*p*-coumaraldehyde	L.E than *p*-coumaryl alcohol
Sinapaldehyde	Partition to membrane	Sinapaldehyde	Partition to membrane
p-coumaryl alcohol	Pool in S.C.W	*p*-coumaryl alcohol	Pool in S.C.W
Coniferyl alcohol	Pool in S.C.W	Coniferyl alcohol	Pool in S.C.W
Sinapyl alcohol	Pool in S.C.W	Sinapyl alcohol	Pool in S.C.W

L.E. is label enrichment
S.C.W is secondary cell wall

A3. Targeted Metabolomics of the Phenylpropanoid Pathway in Arabidopsis Mutants

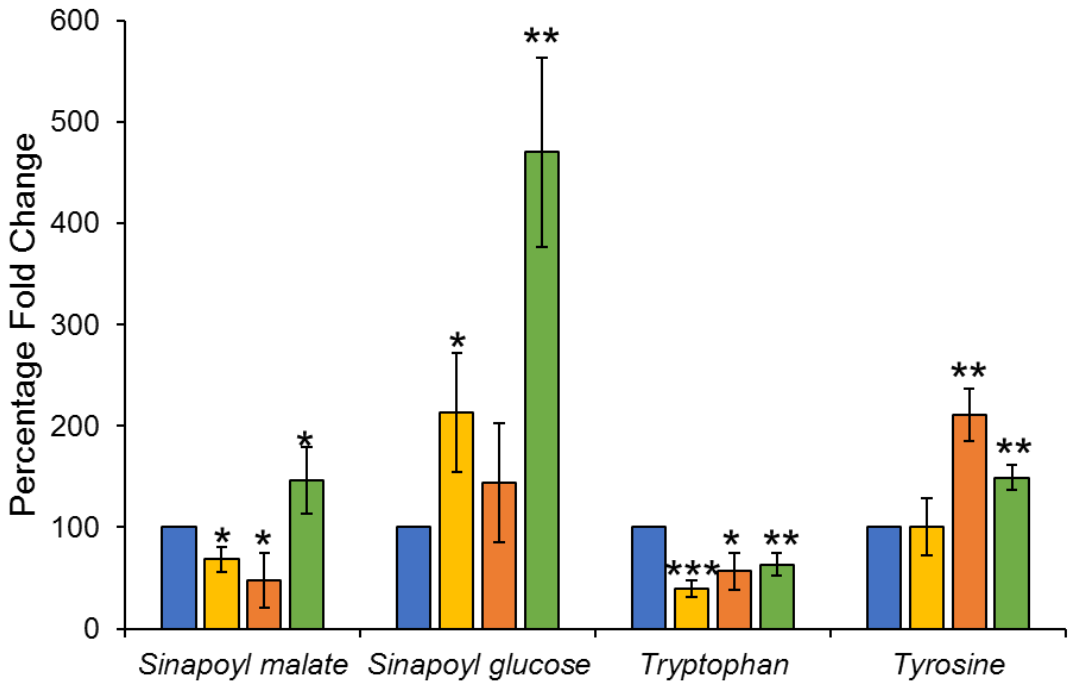


Figure A3.1: Relative abundances of phenylpropanoid intermediates and amino acids in 5 week old whole stems of Arabidopsis WT (blue), *cse2* (yellow), *med5a/5b ref8-1* (orange), and *ccr1* (green) lines. Data presented as mean ± S.D. from n=3 replicates. *p < 0.05, **p < 0.01, and ***p < 0.001 were obtained using standard Student's *t*-test. Peak areas were normalized to WT plants.

A4. Machine Learning Driven Estimation of Optimal Lignin Content and Composition

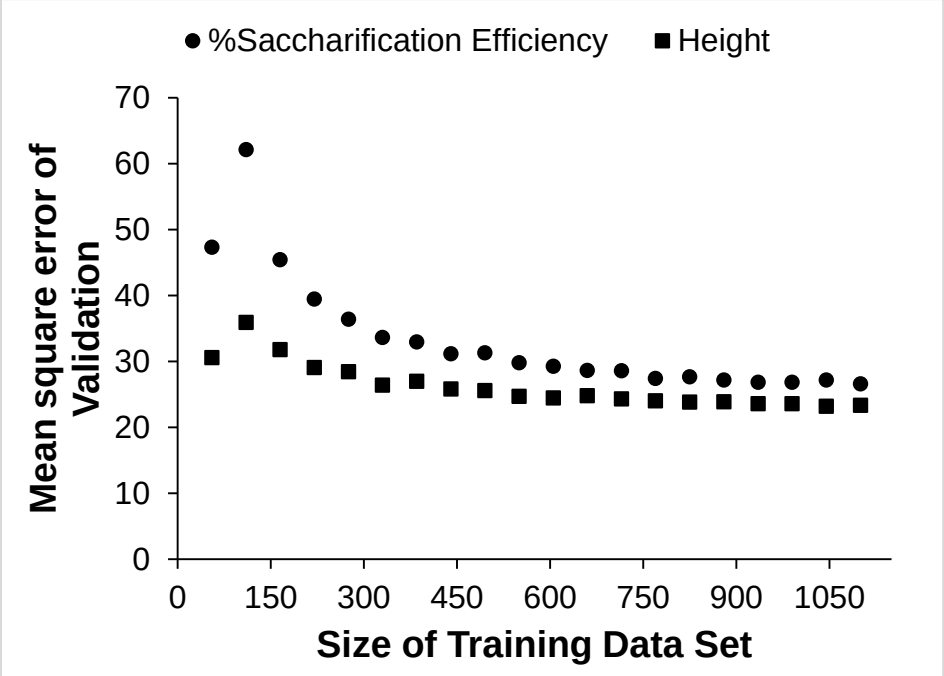


Figure A4.1: Validation performance of SVR models as a function of the training data set size. SVR models were trained to predict the target variables %saccharification efficiency (circle) and plant height (square).

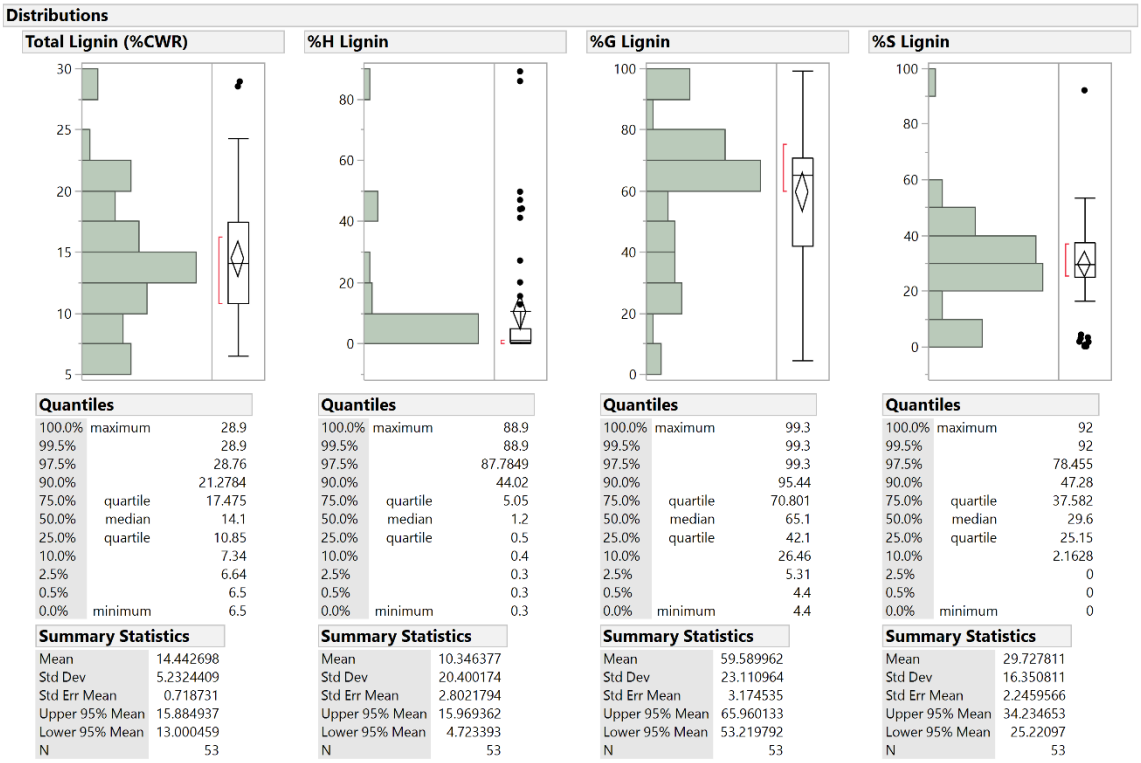


Figure A4.2: Distribution of the input variables in the original data set of 53 lines. The numbers on the bars in the distribution histograms represent the counts.

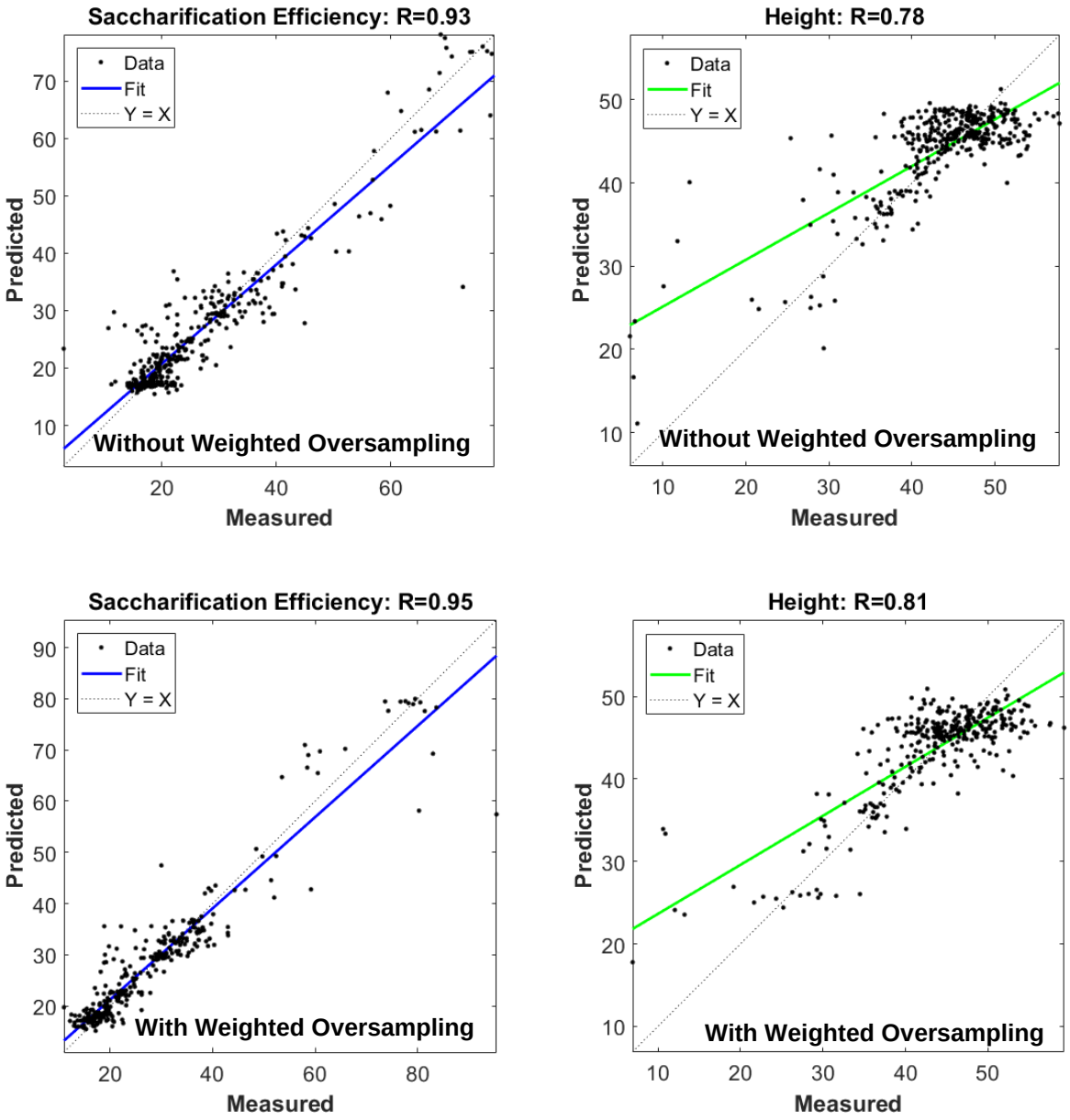


Figure A4.3: Performance of the SVR models on validation data with and without oversampling of the underrepresented lines.

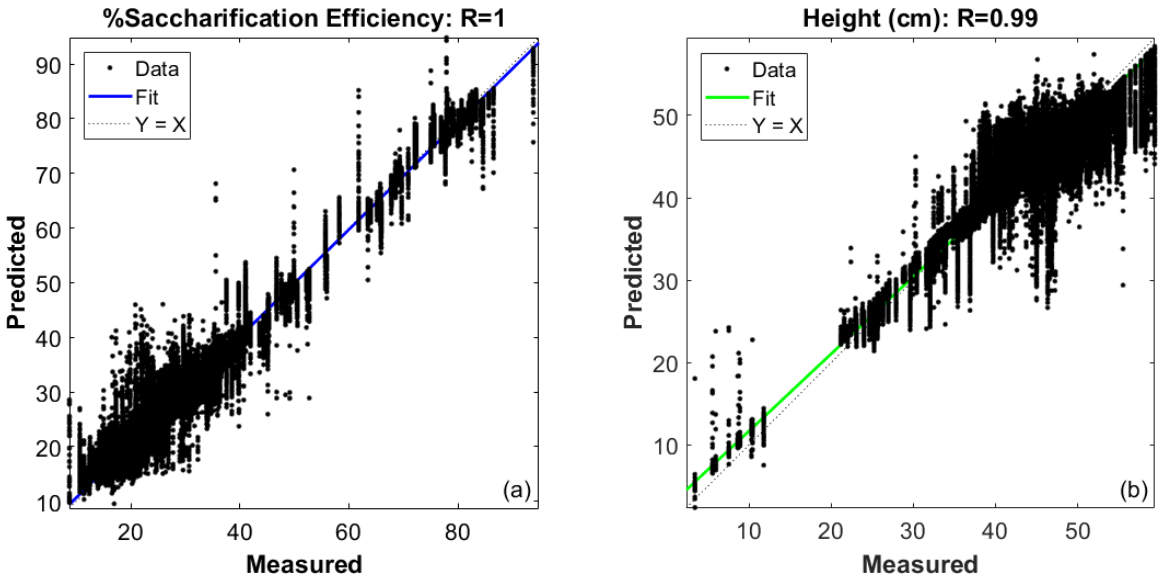


Figure A4.4: Performance of the SVR models on training data after optimizing the hyperparameters.

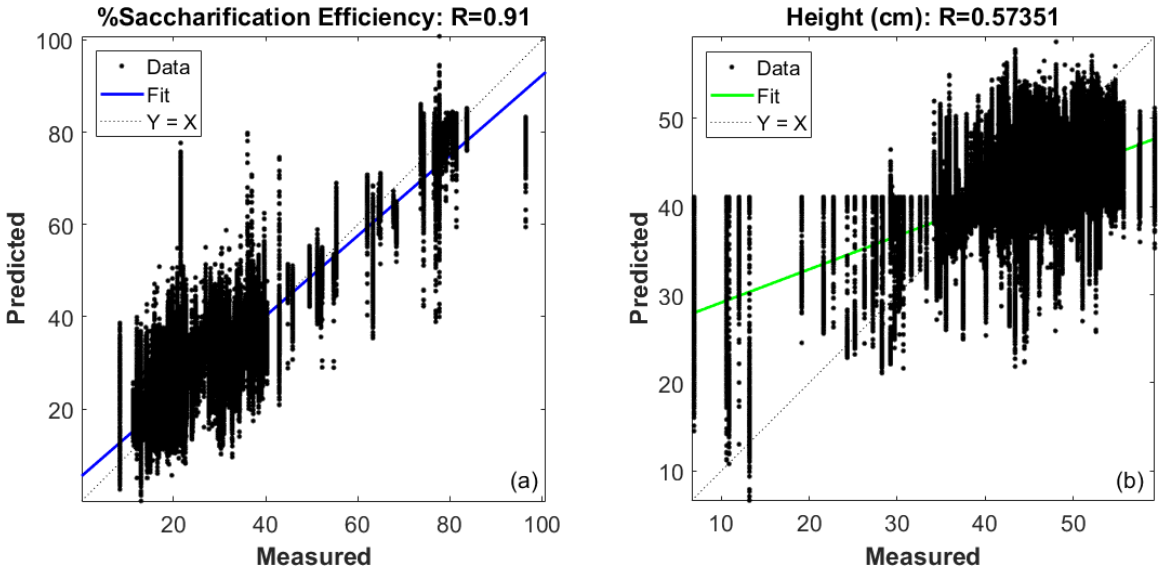


Figure A4.5: Performance of the SVR models with optimized hyperparameters on validation data.

A5. Non-aqueous fractionation of Arabidopsis Stems

Table A5.1: Linear density gradient used for the NAQF procedure on Arabidopsis stems.

Fraction top(sample)	**Density (g/cm^3) 1.3**	**ml required 5**	**ml made 5 (15)**	**ml of 1.3 5 (15)**	**ml of 1.35**	**ml of 1.4**	**ml of 1.5**	**ml of 1.6**
2	1.32	2	3 (9)	1.8 (5.4)	1.2 (3.6)			
3	1.34	2	3 (9)	0.6 (1.8)	2.4 (7.2)			
4	1.36	2	3 (9)	2.4 (7.2)	0.6 (1.8)			
5	1.38	2	3 (9)		1.2 (3.6)	1.8 (5.4)		
6	1.40	2	3 (9)		0	3 (9)		
7	1.42	2	3 (9)		2.4 (7.2)	0.6 (1.8)		
8	1.44	2	3 (9)		1.8 (5.4)	1.2 (3.6)		
9	1.46	2	3 (9)			1.2 (3.6)	1.8 (5.4)	
10	1.48	2	3 (9)			0.6 (1.8)	2.4 (7.2)	
11	1.50	2	3 (9)			0	3 (9)	
12	1.52	2	3 (9)			2.4 (7.2)	0.6 (1.8)	
13	1.54	2	3 (9)			1.8 (5.4)	1.2 (3.6)	
14	1.56	2	3 (9)			1.2 (3.6)	1.8 (5.4)	
15	1.58	2	3 (9)				0.6 (1.8)	2.4 (7.2)
16	1.6	2	3 (9)				0	3 (9)
bottom	1.62	5	5 (15)				TetraCE from bottle	
Total Volume (ml)		**40**		**9.8 (29.4)**	**9.6 (28.8)**	**13.8 (41.4)**	**11.4 (34.2)**	**5.4 (16.2)**

APPENDIX B: PROTOCOLS

B1. Buffer recipes for conducting NAQF

The sections below show the recipes for enzyme extraction buffer (EB) and other buffers required for enzyme assays. The first four buffers are used either for master mixes (MM). If the buffers have been left in the fridge for several weeks, it is wise to verify that the buffer pH remains unchanged.

B1.1 110 mM Tris-Sulfate Buffer (pH 8.5; 100 ml)

- Add 1.332 g Trizma base to 91.5 ml ultra-pure H_2O and mix until complete dissolution.
- Add ~610.7 µl H_2SO_4.
- Add 8.5 ml of 2M NaOH to the solution mixture.
- Measure pH after mixing. Add NaOH to adjust pH if required.
- Filter buffer solution using Millipore 0.22 µm filter into a 250ml autoclaved flask.
- Label and store for future use at 2-8 ^{0}C.

B1.2 100 mM HEPES-$MgCl_2$ buffer (pH 7.8; 100 ml)

- Add 2.383 g of HEPES and 203 mg of $MgCl_2.6H_2O$ to 96.1ml of ultra-pure H_2O and mix until complete dissolution.
- Add 3.9 ml of 2M KOH and mix well.
- Measure pH after mixing. Add KOH to adjust pH if required.
- Filter buffer solution using Millipore 0.22 µm filter into a glass bottle.
- Label and store at 2-8 ^{0}C.

B1.3 100 mM Citric Acid Buffer (pH 4.5; 100 ml)

- Add 1.92 g of citric acid to 91.25 ml of ultra-pure H_2O and mix until complete dissolution.
- Add 8.75 ml of 2M NaOH to the solutions.
- Measure pH after mixing. Add NaOH to adjust pH if required.

- Filter buffer solution using Millipore 0.22 μm filter into a glass bottle.
- Label and store for future use at 2-8 ^{0}C.

B1.4 200 mM Borate Buffer (pH 9.8, 100 ml)

- Add 1.24 g of boric acid to 92.75 ml of ultra-pure H_2O and mix until complete dissolution.
- Add 7.25 ml of 2M NaOH to the solutions.
- Measure pH after mixing. Add NaOH to adjust pH if required.
- Filter buffer solution using Millipore 0.22 μm filter into a glass bottle.
- Label and store for future use at 2-8 ^{0}C.

B1.5 Enzyme extraction buffer EB (pH 7.8, 100ml):

- Dilute 50ml HEPES-$MgCl_2$ buffer to 100ml using ultra-pure water, then add the ingredients according to the table below.
- Mix well. Check for pH and filter the buffer using Millipore 0.22 μm filter into a glass bottle.
- Label and store for future use at 2-8 ^{0}C.

Table B1.1: List of chemicals used for preparing the extraction buffer.

Chemical	Required Cocnentration (mM)	Molecular Weight (g/mol)	Weight to be added (mg)	Notes	Location in FRNY 1190
EDTA	1	292.4	29.24		Aisle 1
PVPP (g/l)	2		200	Add only 2-3 days before experiment as it reduces shelf life of EB.	Aisle 2
DTT	1	154.3	15.43		-20°C freezer
e-aminocaproic acid	1	131.2	13.12		Aisle 2
benzamidine	1	120.15	12.015		Aisle 2
PMSF	1.5	174.2	26.13	Has a half-life of 30 minutes at pH 8 in aqueous solutions. Add before being used.	Aisle 2
Triton X-100 (%v/v)	0.1		100 ul		Aisle 2

EDTA, Ethylenediaminetetraacetic acid; PVPP, Polyvinylpolypyrrolidone; DTT, Dithiothreitol; PMSF, Phenylmethanesulfonyl fluoride.

B.2 Preparation of master mixes (MM) for marker enzyme assays.

The three different assays conducted were; ADP-Glucose pyrophosphorylase (AGPase) assay as plastidial marker, Phosphoenolpyruvate (PEP) as a cytosolic marker, and *a*-mannosidase as a vacuolar marker. MM preparation for each assay have been listed in the following tables.

Table B2.1 AGPase (Plastidial Assay)

Solution/Chemical	Full name	Stock conc [mM]	Vol added to MM [µL]	Notes
HEPES-MgCl2	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid and Magnesium Chloride mixture (pH 7.8)	100	4200	Use HEPES buffer prepared earlier
PGA (aliquots)	D-(-)-3-Phosphoglyceric acid disodium salt	300	84	Aliquoted to correct concentration
DTT	DL-Dithiothreitol	300	84	Solution made on day of assay
β-NADP (aliquots)	β-Nicotinamide adenine dinucleotide phosphate	10	420	Aliquoted to correct concentration

B2.1 continued

ADP Glucose (aliquots)	Adenosine-5′ diphosphoglucose disodium salt	10	840	Aliquoted to correct concentration
G16 (aliquots)	Glucose 1,6-diphosphate	1	84	Aliquoted to correct concentration
PGM	Phosphoglucomutase from rabbit muscle	885 U/ml	15.12	Add only the U/ml required. Found in-20°C
G6PDH	Glucose-6-Phosphate Dehydrogenase	250 U/ml	23.52	
UP H2O			2649.36	
Total			8400	
Substrate				
NAPPi	Sodium Pyrophosphate	25	Not in MM	Found in Aisle 2.

Table B2.2 PEP-Carboxylase (Cytosolic Assay)

Chemical/Solition	Full name	Stock conc [mM]	Vol added to MM[µL]	Notes
Tris-Sulfate buffer		110	6120	Prepared before.
MgSO4	Magnesium Sulfate	300	510	
β-NADH (aliquots)	β-Nicotinamide adenine dinucleotide, reduced dipotassium salt	6	510	
NaHCO3	Sodium Bicarbonate	100	1530	Prepared on day
1,4-Dioxane	-		1530	Flammable cupboard
DTE	Dithioerythritol	300	510	Prepared on day
MDH	Malate dehydrogenase	600U/mL protein	51	Found in-20°C
Substrate				
PEP	Phosphoenol pyruvate	30	Not in MM	Found in-20°C

Table B2.3 α-mannosidase assay (Vacuolar Marker)

	Full name	Stock conc [mM]	Per assay [µL]	Notes
Buffer				
Citrate Buffer	-	100	44	Made previously.
Substrate				
4PNP	4-Nitrophenyl α-D-mannopyranoside	21.85	40	Found in -20°C
Stopping Buffer				
Borate Buffer	-	200	0	Made previously.

4PNP, *p*-nitrophenol pyrranoside

REFERENCES

1. Saini JK, Saini R, Tewari L. Lignocellulosic agriculture wastes as biomass feedstocks for second-generation bioethanol production: concepts and recent developments. 3 Biotech. 2015;5: 337–353. doi:10.1007/s13205-014-0246-5
2. Limayem A, Ricke SC. Lignocellulosic biomass for bioethanol production: Current perspectives, potential issues and future prospects. Prog Energy Combust Sci. Elsevier Ltd; 2012;38: 449–467. doi:10.1016/j.pecs.2012.03.002
3. Sánchez ÓJ, Cardona CA. Trends in biotechnological production of fuel ethanol from different feedstocks. Bioresour Technol. 2008;99: 5270–5295. doi:10.1016/j.biortech.2007.11.013
4. Zabed H, Sahu JN, Boyce AN, Faruq G. Fuel ethanol production from lignocellulosic biomass: An overview on feedstocks and technological approaches. Renew Sustain Energy Rev. Elsevier; 2016;66: 751–774. doi:10.1016/j.rser.2016.08.038
5. Zabed H, Sahu JN, Suely A, Boyce AN, Faruq G. Bioethanol production from renewable sources: Current perspectives and technological progress. Renew Sustain Energy Rev. Elsevier Ltd; 2017;71: 475–501. doi:10.1016/j.rser.2016.12.076
6. Vanholme R, Morreel K, Ralph J, Boerjan W. Lignin engineering. Curr Opin Plant Biol. 2008;11: 278–85. doi:10.1016/j.pbi.2008.03.005
7. Chapple C, Ladisch M, Meilan R. Loosening lignin's grip on biofuel production. Nat Biotechnol. 2007;25: 746–8. doi:10.1038/nbt0707-746
8. Loqu?? D, Scheller H V., Pauly M. Engineering of plant cell walls for enhanced biofuel production. Curr Opin Plant Biol. 2015;25: 151–161. doi:10.1016/j.pbi.2015.05.018
9. Mosier N, Wyman C, Dale B, Elander R, Lee YY, Holtzapple M, et al. Features of promising technologies for pretreatment of lignocellulosic biomass. Bioresour Technol. 2005;96: 673–686. doi:10.1016/j.biortech.2004.06.025

10. Hill J, Nelson E, Tilman D, Polasky S, Tiffany D. Environmental, economic, and energetic costs and benefits of biodiesel and ethanol biofuels. Proc Natl Acad Sci. 2006;103: 11206–11210. doi:10.1073/pnas.0604600103
11. da Costa Sousa L, Chundawat SP, Balan V, Dale BE. "Cradle-to-grave" assessment of existing lignocellulose pretreatment technologies. Curr Opin Biotechnol. 2009;20: 339–347. doi:10.1016/j.copbio.2009.05.003
12. Axelsson L, Franzén M, Ostwald M, Berndes G, Lakshmi G, Ravindranath NH. Perspective: Jatropha cultivation in southern India: Assessing farmers' experiences. Biofuels, Bioprod Biorefining. 2012;6: 246–256. doi:10.1002/bbb
13. Li K, Qin JC, Liu CG, Bai FW. Optimization of pretreatment, enzymatic hydrolysis and fermentation for more efficient ethanol production by Jerusalem artichoke stalk. Bioresour Technol. Elsevier Ltd; 2016;221: 188–194. doi:10.1016/j.biortech.2016.09.021
14. Papa G, Varanasi P, Sun L, Cheng G, Stavila V, Holmes B, et al. Exploring the effect of different plant lignin content and composition on ionic liquid pretreatment efficiency and enzymatic saccharification of Eucalyptus globulus L. mutants. Bioresour Technol. Elsevier Ltd; 2012;117: 352–359. doi:10.1016/j.biortech.2012.04.065
15. Boerjan W, Meyermans H, Chen C, Baucher M, Doorsselaere J Van, Morreel K, et al. Lignin biosynthesis in poplar: Genetic engineering and effects on kraft pulping. Prog Biotechnol. 2001;18: 187–194. doi:10.1016/S0921-0423(01)80072-1
16. Li X, Weng JK, Chapple C. Improvement of biomass through lignin modification. Plant J. 2008;54: 569–581. doi:10.1111/j.1365-313X.2008.03457.x
17. Weng J-K, Li X, Bonawitz ND, Chapple C. Emerging strategies of lignin engineering and degradation for cellulosic biofuel production. Curr Opin Biotechnol. 2008;19: 166–72. doi:10.1016/j.copbio.2008.02.014
18. Welker CM, Balasubramanian VK, Petti C, Rai KM, De Bolt S, Mendu V. Engineering plant biomass lignin content and composition for biofuels and bioproducts. Energies. 2015;8: 7654–7676. doi:10.3390/en8087654
19. Fraser CM, Chapple C. The phenylpropanoid pathway in Arabidopsis. Arabidopsis Book. 2011;9: e0152. doi:10.1199/tab.0152

20. Vogt T. Phenylpropanoid biosynthesis. Mol Plant. 2010;3: 2–20. doi:10.1093/mp/ssp106
21. Famili I, Forster J, Nielsen J, Palsson BO. Saccharomyces cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. Proc Natl Acad Sci U S A. 2003;100: 13134–9. doi:10.1073/pnas.2235812100
22. Li Y, Kim JI, Pysh L, Chapple C. Four isoforms of Arabidopsis thaliana 4-coumarate: CoA ligase (4CL) have overlapping yet distinct roles in phenylpropanoid metabolism. Plant Physiol. 2015;169: pp.00838.2015. doi:10.1104/pp.15.00838
23. Hoffmann L, Besseau S, Geoffroy P, Ritzenthaler C, Meyer D, Lapierre C, et al. Silencing of hydroxycinnamoyl-coenzyme A shikimate/quinate hydroxycinnamoyltransferase affects phenylpropanoid biosynthesis. Plant Cell. 2004;16: 1446–65. doi:10.1105/tpc.020297
24. Hoffmann L, Maury S, Martz F, Geoffroy P, Legrand M. Purification, cloning, and properties of an acyltransferase controlling shikimate and quinate ester intermediates in phenylpropanoid metabolism. J Biol Chem. 2003;278: 95–103. doi:10.1074/jbc.M209362200
25. Vanholme R, Cesarino I, Rataj K, Xiao Y, Sundin L, Goeminne G, et al. Caffeoyl Shikimate Esterase (CSE) Is an Enzyme in the Lignin Biosynthetic Pathway in Arabidopsis. Science. 2013: 1103–1107.
26. Wang P, Dudareva N, Morgan JA, Chapple C. Genetic manipulation of lignocellulosic biomass for bioenergy. Curr Opin Chem Biol. Elsevier Ltd; 2015;29: 32–39. doi:10.1016/j.cbpa.2015.08.006
27. Li Q, Song J, Peng S, Wang JP, Qu GZ, Sederoff RR, et al. Plant biotechnology for lignocellulosic biofuel production. Plant Biotechnol J. 2014;12: 1174–1192. doi:10.1111/pbi.12273
28. Bonawitz ND, Chapple C. The genetics of lignin biosynthesis: connecting genotype to phenotype. Annu Rev Genet. 2010;44: 337–63. doi:10.1146/annurev-genet-102209-163508

29. Besseau S, Hoffmann L, Geoffroy P, Lapierre C, Pollet B, Legrand M. Flavonoid Accumulation in Arabidopsis Repressed in Lignin Synthesis Affects Auxin Transport and Plant Growth. Plant Cell Online. 2007;19: 148–162. doi:10.1105/tpc.106.044495
30. Zhao Q, Tobimatsu Y, Zhou R, Pattathil S, Gallego-giraldo L, Fu C. Loss of function of cinnamyl alcohol dehydrogenase 1 leads to unconventional lignin and a temperature- sensitive growth defect in Medicago truncatula. Proc Natl Acad Sci U S A. 2013;110: 13660–13665. doi:10.1073/pnas.1312234110/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1312234110
31. Lu F, Marita JM, Lapierre C, Jouanin L, Morreel K, Boerjan W, et al. Sequencing around 5-hydroxyconiferyl alcohol-derived units in caffeic acid O-methyltransferase-deficient poplar lignins. Plant Physiol. 2010;153: 569–79. doi:10.1104/pp.110.154278
32. Weng JK, Mo H, Chapple C. Over-expression of F5H in COMT-deficient Arabidopsis leads to enrichment of an unusual lignin and disruption of pollen wall formation. Plant J. 2010;64: 898–911. doi:10.1111/j.1365-313X.2010.04391.x
33. Chen F, Dixon RA. Lignin modification improves fermentable sugar yields for biofuel production. Nat Biotechnol. 2007;25: 759–761. doi:10.1038/nbt1316
34. Liu B, G??mez LD, Hua C, Sun L, Ali I, Huang L, et al. Linkage mapping of stem saccharification digestibility in rice. PLoS One. 2016;11: 1–13. doi:10.1371/journal.pone.0159117
35. Van Acker R, Vanholme R, Storme V, Mortimer JC, Dupree P, Boerjan W. Lignin biosynthesis perturbations affect secondary cell wall composition and saccharification yield in Arabidopsis thaliana. Biotechnol Biofuels. 2013;6: 46. doi:10.1186/1754-6834-6-46
36. Vanholme R, Demedts B, Morreel K, Ralph J, Boerjan W. Lignin biosynthesis and structure. Plant Physiol. 2010;153: 895–905. doi:10.1104/pp.110.155119
37. Hood EE. Plant-based biofuels. F1000Research. 2016;5: 1–9. doi:10.12688/f1000research.7418.1

38. Bonawitz ND, Chapple C. The genetics of lignin biosynthesis: connecting genotype to phenotype. Annu Rev Genet. 2010;44: 337–63. doi:10.1146/annurev-genet-102209-163508
39. Shi R, Sun Y-H, Li Q, Heber S, Sederoff R, Chiang VL. Towards a systems approach for lignin biosynthesis in Populus trichocarpa: transcript abundance and specificity of the monolignol biosynthetic genes. Plant Cell Physiol. 2010;51: 144–63. doi:10.1093/pcp/pcp175
40. Lee Y, Chen F, Gallego-Giraldo L, Dixon R a, Voit EO. Integrative analysis of transgenic alfalfa (Medicago sativa L.) suggests new metabolic control mechanisms for monolignol biosynthesis. PLoS Comput Biol. 2011;7: e1002047. doi:10.1371/journal.pcbi.1002047
41. Lee Y, Escamilla-Treviño L, Dixon R a, Voit EO. Functional analysis of metabolic channeling and regulation in lignin biosynthesis: a computational approach. PLoS Comput Biol. 2012;8: e1002769. doi:10.1371/journal.pcbi.1002769
42. Moseley HNB. Error Analysis and Propagation in Metabolomics Data Analysis. Comput Struct Biotechnol J. Research Network of Computational and Structural Biotechnology; 2013;4: 1–12. doi:10.5936/csbj.201301006
43. Noack S, Wiechert W. Quantitative metabolomics: a phantom? Trends Biotechnol. Elsevier Ltd; 2014;32: 238–244. doi:10.1016/j.tibtech.2014.03.006
44. Kueger S, Steinhauser D, Willmitzer L, Giavalisco P. High-resolution plant metabolomics: from mass spectral features to metabolites and from whole-cell analysis to subcellular metabolite distributions. Plant J. 2012;70: 39–50. doi:10.1111/j.1365-313X.2012.04902.x
45. Sumner LW, Lei Z, Nikolau BJ, Saito K. Modern plant metabolomics: advanced natural product gene discoveries, improved technologies, and future prospects. Nat Prod Rep. Royal Society of Chemistry; 2015;32: 212–29. doi:10.1039/c4np00072b
46. Rochfort SJ, Trenerry VC, Imsic M, Panozzo J, Jones R. Class targeted metabolomics: ESI ion trap screening methods for glucosinolates based on MSn fragmentation. Phytochemistry. 2008;69: 1671–9. doi:10.1016/j.phytochem.2008.02.010

47. Hegeman AD. Plant metabolomics--meeting the analytical challenges of comprehensive metabolite analysis. Brief Funct Genomics. 2010;9: 139–48. doi:10.1093/bfgp/elp053

48. de la Torre-Carbot K, Jauregui O, Castellote AI, Lamuela-Raventós RM, Covas M-I, Casals I, et al. Rapid high-performance liquid chromatography-electrospray ionization tandem mass spectrometry method for qualitative and quantitative analysis of virgin olive oil phenolic metabolites in human low-density lipoproteins. J Chromatogr A. 2006;1116: 69–75. doi:10.1016/j.chroma.2006.03.022

49. Allwood JW, Ellis DI, Goodacre R. Metabolomic technologies and their application to the study of plants and plant-host interactions. Physiol Plant. 2008;132: 117–135. doi:10.1111/j.1399-3054.2007.01001.x

50. Arrivault S, Guenther M, Ivakov A, Feil R, Vosloh D, van Dongen JT, et al. Use of reverse-phase liquid chromatography, linked to tandem mass spectrometry, to profile the Calvin cycle and other metabolic intermediates in Arabidopsis rosettes at different carbon dioxide concentrations. Plant J. 2009;59: 826–39. doi:10.1111/j.1365-313X.2009.03902.x

51. Meyermans H, Morreel K, Lapierre C, Pollet B, De Bruyn a, Busson R, et al. Modifications in lignin and accumulation of phenolic glucosides in poplar xylem upon down-regulation of caffeoyl-coenzyme A O-methyltransferase, an enzyme involved in lignin biosynthesis. J Biol Chem. 2000;275: 36899–909. doi:10.1074/jbc.M006915200

52. Chen F, Duran AL, Blount JW, Sumner LW, Dixon RA. Profiling phenolic metabolites in transgenic alfalfa modified in lignin biosynthesis. Phytochemistry. 2003;64: 1013–1021. doi:10.1016/S0031-9422(03)00463-1

53. Damiani I, Morreel K, Danoun S, Goeminne G, Yahiaoui N, Marque C, et al. Metabolite profiling reveals a role for atypical cinnamyl alcohol dehydrogenase CAD1 in the synthesis of coniferyl alcohol in tobacco xylem. Plant Mol Biol. 2005;59: 753–69. doi:10.1007/s11103-005-0947-6

54. Long M, Millar DJ, Kimura Y, Donovan G, Rees J, Fraser PD, et al. Metabolite profiling of carotenoid and phenolic pathways in mutant and transgenic lines of tomato: Identification of a high antioxidant fruit line. Phytochemistry. 2006;67: 1750–1757. doi:10.1016/j.phytochem.2006.02.022
55. Lin LZ, Harnly JM, Upton R. Comparison of the phenolic component profiles of skullcap (Scutellaria lateriflora) and germander (Teucrium canadense and T. chamaedrys), a potentially hepatotoxic adulterant. Phytochem Anal. 2009;20: 298–306. doi:10.1002/pca.1127
56. Callipo L, Cavaliere C, Fuscoletti V, Gubbiotti R, Samperi R, Lagan?? A. Phenilpropanoate identification in young wheat plants by liquid chromatography/tandem mass spectrometry: Monomeric and dimeric compounds. J Mass Spectrom. 2010;45: 1026–1040. doi:10.1002/jms.1800
57. Liu J, Shi R, Li Q, Sederoff RR, Chiang VL. A standard reaction condition and a single HPLC separation system are sufficient for estimation of monolignol biosynthetic pathway enzyme activities. Planta. 2012;236: 879–85. doi:10.1007/s00425-012-1688-9
58. Owen BC, Haupert LJ, Jarrell M, Marcum CL, Parsell TH, Abu-omar MM, et al. High-Performance Liquid Chromatography/High-Resolution Multiple Stage Tandem Mass Spectrometry Using Negative-Ion-Mode Hydroxide-Doped Electrospray Ionization for the Characterization of Lignin Degradation Products. Anal Chem. 2012;84: 6000–6007.
59. Frolov A, Henning A, Bo C, Tissier A, Strack D. An UPLC-MS/MS Method for the Simultaneous Identification and Quantitation of Cell Wall Phenolics in Brassica napus Seeds. J Agric Food Chem. 2013;61: 1219–1227.
60. Reuben S, Rai A, Pillai BVS, Rodrigues A, Swarup S. A bacterial quercetin oxidoreductase quoa-mediated perturbation in the phenylpropanoid metabolic network increases lignification with a concomitant decrease in phenolamides in arabidopsis. J Exp Bot. 2013;64: 5183–5194. doi:10.1093/jxb/ert310
61. Ferreres F, Oliveira AP, Gil-Izquierdo A, Valentão P, Andrade PB. Piper betle leaves: Profiling phenolic compounds by HPLC/DAD-ESI/MS n and anti-cholinesterase activity. Phytochem Anal. 2014;25: 453–460. doi:10.1002/pca.2515

62. Shao Y, Jiang J, Ran L, Lu C, Wei C, Wang Y. Analysis of flavonoids and hydroxycinnamic acid derivatives in rapeseeds (Brassica napus L. var. napus) by HPLC-PDA--ESI(--)-MS(n)/HRMS. J Agric Food Chem. 2014;62: 2935–45. doi:10.1021/jf404826u

63. Eudes A, Sathitsuksanoh N, Baidoo EEK, George A, Liang Y, Yang F, et al. Expression of a bacterial 3-dehydroshikimate dehydratase reduces lignin content and improves biomass saccharification efficiency. Plant Biotechnol J. 2015;13: 1241–1250. doi:10.1111/pbi.12310

64. Mocan A, Schafberg M, Crișan G, Rohn S. Determination of lignans and phenolic components of Schisandra chinensis (Turcz.) Baill. using HPLC-ESI-ToF-MS and HPLC-online TEAC: Contribution of individual components to overall antioxidant activity and comparison with traditional antioxidant assays. J Funct Foods. 2016;24: 579–594. doi:10.1016/j.jff.2016.05.007

65. Šibul F, Orčić D, Vasić M, Anačkov G, Nadpal J, Savić A, et al. Phenolic profile, antioxidant and anti-inflammatory potential of herb and root extracts of seven selected legumes. Ind Crops Prod. 2016;83: 641–653. doi:10.1016/j.indcrop.2015.12.057

66. ICH. ICH Topic Q2 (R1) Validation of Analytical Procedures : Text and Methodology. Int Conf Harmon. 2005;1994: 17.

67. Bajad SU, Lu W, Kimball EH, Yuan J, Peterson C, Rabinowitz JD. Separation and quantitation of water soluble cellular metabolites by hydrophilic interaction chromatography-tandem mass spectrometry. J Chromatogr A. 2006;1125: 76–88. doi:10.1016/j.chroma.2006.05.019

68. Giorgianni F, Cappiello A, Beranova-giorgianni S, Palma P, Trufelli H, Desiderio DM. LC - MS / MS Analysis of Peptides with Methanol as Organic Modifier : Improved Limits of Detection. Anal Chem. 2004;76: 7028–7038.

69. Hashem H, Tründelberg C, Attef O, Jira T. Effect of chromatographic conditions on liquid chromatographic chiral separation of terbutaline and salbutamol on Chirobiotic V column. J Chromatogr A. Elsevier B.V.; 2011;1218: 6727–31. doi:10.1016/j.chroma.2011.07.090

70. Kamel AM, Brown PR, Munson B. Effects of Mobile-Phase Additives, Solution pH, Ionization Constant, and Analyte Concentration on the Sensitivities and Electrospray Ionization Mass Spectra of Nucleoside Antiviral Agents. Anal Chem. 1999;71: 5481–5492. doi:10.1021/ac9906429
71. Liigand J, Kruve A, Leito I, Girod M, Antoine R. Effect of mobile phase on electrospray ionization efficiency. J Am Soc Mass Spectrom. 2014;25: 1853–61. doi:10.1007/s13361-014-0969-x
72. Lloyd SR, Kirkland JJ, Dolan JW. Introduction to Modern Liquid Chromatography. 2010.
73. Hua Y, Jenke D, Corporation BH, Division TR, Route W, Lake R. Increasing the Sensitivity of an LC – MS Method for Screening Material Extracts for Organic Extractables via Mobile Phase Optimization. J Chromatogr Sci. 2012;50: 213–227.
74. Wu Z, Gao W, Phelps MA, Wu D, Miller DD, James T. Favorable Effects of Weak Acids on Negative-Ion Electrospray Ionization Mass Spectrometry. Anal Biochem. 2004;76: 839–847.
75. Constantopoulos TL, Jackson GS, Enke CG. Effects of Salt Concentration on Analyte Response Using Electrospray Ionization Mass Spectrometry. J Am Soc Mass Spectrom. 1999;10: 625–634.
76. Cech NB, Enke CG. Selectivity in Electrospray Ionization Mass Spectrometry. Electrospray and MALDI Mass Spectrometry: Fundamentals, Instrumentation, Practicalities, and Biological Applications. 2010. pp. 49–73.
77. Yang Y, Lamm LJ, He P, Kondo T. Temperature Effect on Peak Width and Column Efficiency in Subcritical Water Chromatography. J Chromatogr Sci. 2002;40: 107–112.
78. Shi S, Valle-Rodríguez JO, Khoomrung S, Siewers V, Nielsen J. Functional expression and characterization of five wax ester synthases in Saccharomyces cerevisiae and their utility for biodiesel production. Biotechnol Biofuels. BioMed Central Ltd; 2012;5: 7. doi:10.1186/1754-6834-5-7

79. Dent M, Dragović-Uzelac V, Penić M, Brñić M, Bosiljkov T, Levaj B. The effect of extraction solvents, temperature and time on the composition and mass fraction of polyphenols in dalmatian wild sage (Salvia officinalis L.) extracts. Food Technol Biotechnol. 2013;51: 84–91.
80. Khoddami A, Wilkes M a., Roberts TH. Techniques for analysis of plant phenolic compounds. Molecules. 2013;18: 2328–2375. doi:10.3390/molecules18022328
81. Römisch-Margl W, Prehn C, Bogumil R, Röhring C, Suhre K, Adamski J. Procedure for tissue sample preparation and metabolite extraction for high-throughput targeted metabolomics. Metabolomics. 2012;8: 133–142. doi:10.1007/s11306-011-0293-4
82. Hall RD. Plant metabolomics: from holistic hope, to hype, to hot topic. New Phytol. 2006;169: 453–68. doi:10.1111/j.1469-8137.2005.01632.x
83. De Vos RCH, Moco S, Lommen A, Keurentjes JJB, Bino RJ, Hall RD. Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry. Nat Protoc. 2007;2: 778–91. doi:10.1038/nprot.2007.95
84. Duportet X, Aggio RBM, Carneiro S, Villas-Bôas SG. The biological interpretation of metabolomic data can be misled by the extraction method used. Metabolomics. 2011;8: 410–421. doi:10.1007/s11306-011-0324-1
85. t'Kindt R, De Veylder L, Storme M, Deforce D, Van Bocxlaer J. LC-MS metabolic profiling of Arabidopsis thaliana plant leaves and cell cultures: optimization of pre-LC-MS procedure parameters. J Chromatogr B Analyt Technol Biomed Life Sci. 2008;871: 37–43. doi:10.1016/j.jchromb.2008.06.039
86. Gonzales GB, Smagghe G, Raes K, Camp J Van. Combined Alkaline Hydrolysis and Ultrasound-Assisted Extraction for the Release of Nonextractable Phenolics from Cauli fl ower (Brassica oleracea var. botrytis) Waste. J Agric Food Chem. 2014;62: 3371–3376.
87. Taylor PJ. Matrix effects: the Achilles heel of quantitative high-performance liquid chromatography-electrospray-tandem mass spectrometry. Clin Biochem. 2005;38: 328–34. doi:10.1016/j.clinbiochem.2004.11.007

88. Gosetti F, Mazzucco E, Zampieri D, Gennaro MC. Signal suppression/enhancement in high-performance liquid chromatography tandem mass spectrometry. J Chromatogr A. Elsevier B.V.; 2010;1217: 3929–37. doi:10.1016/j.chroma.2009.11.060

89. Yaroshenko D V., Kartsova L a. Matrix effect and methods for its elimination in bioanalytical methods using chromatography-mass spectrometry. J Anal Chem. 2014;69: 311–317. doi:10.1134/S1061934814040133

90. Berg T, Strand DH. ^{13}C labelled internal standards--a solution to minimize ion suppression effects in liquid chromatography-tandem mass spectrometry analyses of drugs in biological samples? J Chromatogr A. Elsevier B.V.; 2011;1218: 9366–74. doi:10.1016/j.chroma.2011.10.081

91. Mathias PC, Hayden J a, Laha TJ, Hoofnagle AN. Evaluation of matrix effects using a spike recovery approach in a dilute-and-inject liquid chromatography-tandem mass spectrometry opioid monitoring assay. Clin Chim Acta. Elsevier B.V.; 2014;437: 38–42. doi:10.1016/j.cca.2014.06.023

92. t'Kindt R, Morreel K, Deforce D, Boerjan W, Van Bocxlaer J. Joint GC-MS and LC-MS platforms for comprehensive plant metabolomics: repeatability and sample pre-treatment. J Chromatogr B Analyt Technol Biomed Life Sci. 2009;877: 3572–80. doi:10.1016/j.jchromb.2009.08.041

93. Antignac J-P, de Wasch K, Monteau F, De Brabander H, Andre F, Le Bizec B. The ion suppression phenomenon in liquid chromatography–mass spectrometry and its consequences in the field of residue analysis. Anal Chim Acta. 2005;529: 129–136. doi:10.1016/j.aca.2004.08.055

94. Niessen WMA, Manini P, Andreoli R. MATRIX EFFECTS IN QUANTITATIVE PESTICIDE ANALYSIS USING LIQUID CHROMATOGRAPHY – MASS SPECTROMETRY. Mass Spectrom Rev. 2006;25: 881–899. doi:10.1002/mas

95. Lauvergeat V, Lacomme C, Lacombe E, Lasserre E, Roby D, Grima-Pettenati J. Two cinnamoyl-CoA reductase (CCR) genes from Arabidopsis thaliana are differentially expressed during development and in response to infection with pathogenic bacteria. Phytochemistry. 2001;57: 1187–1195. doi:10.1016/S0031-9422(01)00053-X

96. Baltas M, Lapeyre C, Bedos-Belval F, Maturano M, Saint-Aguet P, Roussel L, et al. Kinetic and inhibition studies of cinnamoyl-CoA reductase 1 from Arabidopsis thaliana. Plant Physiol Biochem. 2005;43: 746–53. doi:10.1016/j.plaphy.2005.06.003

97. Mir Derikvand M, Sierra JB, Ruel K, Pollet B, Do CT, Thévenin J, et al. Redirection of the phenylpropanoid pathway to feruloyl malate in Arabidopsis mutants deficient for cinnamoyl-CoA reductase 1. Planta. 2008;227: 943–956. doi:10.1007/s00425-007-0669-x

98. Ruel K, Berrio-Sierra J, Derikvand MM, Pollet B, Thévenin J, Lapierre C, et al. Impact of CCR1 silencing on the assembly of lignified secondary walls in Arabidopsis thaliana. New Phytol. 2009;184: 99–113. doi:10.1111/j.1469-8137.2009.02951.x

99. Vanholme R, Storme V, Vanholme B, Sundin L, Christensen JH, Goeminne G, et al. A systems biology view of responses to lignin biosynthesis perturbations in Arabidopsis. Plant Cell. 2012;24: 3506–29. doi:10.1105/tpc.112.102574

100. Moinuddin SG a, Jourdes M, Laskar DD, Ki C, Cardenas CL, Kim K-W, et al. Insights into lignin primary structure and deconstruction from Arabidopsis thaliana COMT (caffeic acid O-methyl transferase) mutant Atomt1. Org Biomol Chem. 2010;8: 3928–46. doi:10.1039/c004817h

101. Humphreys JM, Hemm MR, Chapple C. New routes for lignin biosynthesis defined by biochemical characterization of recombinant ferulate 5-hydroxylase, a multifunctional cytochrome P450-dependent monooxygenase. Proc Natl Acad Sci U S A. 1999;96: 10045–50. doi:10.1073/pnas.96.18.10045

102. Jaini R, Wang P, Dudareva N, Chapple C, Morgan JA. Targeted Metabolomics of the Phenylpropanoid Pathway in Arabidopsis thaliana using Reversed Phase Liquid Chromatography Coupled with Tandem Mass Spectrometry. Phytochem Anal. 2017;28: 267–276. doi:10.1002/pca.2672

103. Tumaney AW, Ohlrogge JB, Pollard M. Acetyl coenzyme A concentrations in plant tissues. J Plant Physiol. 2004;161: 485–488. doi:10.1078/0176-1617-01258

104. Larson TR, Graham IA. A novel technique for the sensitive quantification of acyl CoA esters from plant tissues. Plant J. 2001;25: 115–125. doi:10.1046/j.1365-313X.2001.00929.x

105. Haynes CA, Allegood JC, Sims K, Wang EW, Sullards MC, Merrill AH, et al. Quantitation of fatty acyl-coenzyme As in mammalian cells by liquid chromatography-electrospray ionization tandem mass spectrometry. J Lipid Res. 2008;49: 1113–25. doi:10.1194/jlr.D800001-JLR200

106. Qualley A V, Cooper BR, Dudareva N. Profiling hydroxycinnamoyl-coenzyme A thioesters: unlocking the back door of phenylpropanoid metabolism. Anal Biochem. Elsevier Inc.; 2012;420: 182–4. doi:10.1016/j.ab.2011.09.010

107. Bracher PJ, Snyder PW, Bohall BR, Whitesides GM. The Relative Rates of Thiol-Thioester Exchange and Hydrolysis for Alkyl and Aryl Thioalkanoates in Water. Orig Life Evol Biosph. 2011;41: 399–412. doi:10.1007/s11084-011-9243-4

108. Rautengarten C, Baidoo E, Keasling JD, Scheller HV. A simple method for enzymatic synthesis of unlabeled and radiolabeled hydroxycinnamate-CoA. Bioenergy Res. 2010;3: 115–122. doi:10.1007/s12155-010-9085-3

109. Maeda H, Dudareva N. The shikimate pathway and aromatic amino Acid biosynthesis in plants. Annu Rev Plant Biol. 2012;63: 73–105. doi:10.1146/annurev-arplant-042811-105439

110. Fraser CM, Chapple C. The phenylpropanoid pathway in Arabidopsis. Arabidopsis Book. 2011;9: e0152. doi:10.1199/tab.0152

111. Sticklen M. Plant genetic engineering to improve biomass characteristics for biofuels. Curr Opin Biotechnol. 2006;17: 315–9. doi:10.1016/j.copbio.2006.05.003

112. Xiang Z, Sen SK, Min D, Savithri D, Lu F, Jameel H, et al. Field-Grown Transgenic Hybrid Poplar with Modified Lignin Biosynthesis to Improve Enzymatic Saccharification Efficiency. ACS Sustain Chem Eng. 2017;5: 2407–2414. doi:10.1021/acssuschemeng.6b02740

113. Davisonp BH, Drescherp SR, Tuskan GA, Davis MF, Uan AN, Nghiem P. Variation of S/G Ratio and Lignin Content in a Populus Family Influences the Release of Xylose by Dilute Acid Hydrolysis. Appl Biochem Biotechnol. 2006;129–132: 427–435.

114. Schilmiller AL, Stout J, Weng JK, Humphreys J, Ruegger MO, Chapple C. Mutations in the cinnamate 4-hydroxylase gene impact metabolism, growth and development in Arabidopsis. Plant J. 2009;60: 771–782. doi:10.1111/j.1365-313X.2009.03996.x

115. Franke R, Humphreys JM, Hemm MR, Denault JW, Ruegger MO, Cusumano JC, et al. The Arabidopsis REF8 gene encodes the 3-hydroxylase of phenylpropanoid metabolism. Plant J. 2002;30: 33–45. doi:10.1046/j.1365-313X.2002.01266.x

116. Goujon T, Ferret V, Mila I, Pollet B, Ruel K, Burlat V, et al. Down-regulation of the AtCCR1 gene in Arabidopsis thaliana: effects on phenotype, lignins and cell wall degradability. Planta. 2003;217: 218–228. doi:10.1007/s00425-003-0987-6

117. Do CT, Pollet B, Thévenin J, Sibout R, Denoue D, Barrière Y, et al. Both caffeoyl Coenzyme A 3-O-methyltransferase 1 and caffeic acid O-methyltransferase 1 are involved in redundant functions for lignin, flavonoids and sinapoyl malate biosynthesis in Arabidopsis. Planta. 2007;226: 1117–1129. doi:10.1007/s00425-007-0558-3

118. Humphreys JM, Chapple C. Rewriting the lignin roadmap. Curr Opin Plant Biol. 2002;5: 224–229. doi:10.1016/S1369-5266(02)00257-1

119. Schoch G, Goepfert S, Morant M, Hehn a, Meyer D, Ullmann P, et al. CYP98A3 from Arabidopsis thaliana is a 3'-hydroxylase of phenolic esters, a missing link in the phenylpropanoid pathway. J Biol Chem. 2001;276: 36566–74. doi:10.1074/jbc.M104047200

120. Guo D, Chen F, Inoue K, Blount JW, Dixon R a. Downregulation of caffeic acid 3-O-methyltransferase and caffeoyl CoA 3-O-methyltransferase in transgenic alfalfa. impacts on lignin structure and implications for the biosynthesis of G and S lignin. Plant Cell. 2001;13: 73–88. doi:10.1105/tpc.13.1.73

121. Wiechert W, N??h K. Isotopically non-stationary metabolic flux analysis: Complex yet highly informative. Curr Opin Biotechnol. 2013;24: 979–986. doi:10.1016/j.copbio.2013.03.024

122. Schwender J. Metabolic flux analysis as a tool in metabolic engineering of plants. Curr Opin Biotechnol. 2008;19: 131–137. doi:10.1016/j.copbio.2008.02.006

123. O'Grady J, Schwender J, Shachar-Hill Y, Morgan J a. Metabolic cartography: experimental quantification of metabolic fluxes from isotopic labelling studies. J Exp Bot. 2012;63: 2293–308. doi:10.1093/jxb/ers032
124. Libourel IGL, Shachar-Hill Y. Metabolic flux analysis in plants: from intelligent design to rational engineering. Annu Rev Plant Biol. 2008;59: 625–50. doi:10.1146/annurev.arplant.58.032806.103822
125. Matsuda F, Wakasa K, Miyagawa H. Metabolic flux analysis in plants using dynamic labeling technique: Application to tryptophan biosynthesis in cultured rice cells. Phytochemistry. 2007;68: 2290–2301. doi:10.1016/j.phytochem.2007.03.031
126. Paula Alonso A, Dale VL, Shachar-Hill Y. Understanding fatty acid synthesis in developing maize embryos using metabolic flux analysis. Metab Eng. 2010;12: 488–497. doi:10.1016/j.ymben.2010.04.002
127. Sriram G. Quantification of Compartmented Metabolic Fluxes in Developing Soybean Embryos by Employing Biosynthetically Directed Fractional 13C Labeling, Two-Dimensional [13C, 1H] Nuclear Magnetic Resonance, and Comprehensive Isotopomer Balancing. Plant Physiol. 2004;136: 3043–3057. doi:10.1104/pp.104.050625
128. Hay JO, Shi H, Heinzel N, Hebbelmann I, Rolletschek H, Schwender J. Integration of a constraint-based metabolic model of Brassica napus developing seeds with (13)C-metabolic flux analysis. Front Plant Sci. 2014;5: 724. doi:10.3389/fpls.2014.00724
129. Baxter CJ, Redestig H, Schauer N, Repsilber D, Patil KR, Nielsen J, et al. The Metabolic Response of Heterotrophic Arabidopsis Cells to Oxidative Stress 1 [W]. Plant Physiol. 2007;143: 312–325. doi:10.1104/pp.106.090431
130. Szecowka M, Heise R, Tohge T, Nunes-Nesi A, Vosloh D, Huege J, et al. Metabolic fluxes in an illuminated Arabidopsis rosette. Plant Cell. 2013;25: 694–714. doi:10.1105/tpc.112.106989

131. Masakapalli SK, Bryant FM, Kruger NJ, Ratcliffe RG. The metabolic flux phenotype of heterotrophic Arabidopsis cells reveals a flexible balance between the cytosolic and plastidic contributions to carbohydrate oxidation in response to phosphate limitation. Plant J. 2014;78: 964–977. doi:10.1111/tpj.12522

132. Srivastava AC, Chen F, Ray T, Pattathil S, Peña MJ, Avci U, et al. Loss of function of folylpolyglutamate synthetase 1 reduces lignin content and improves cell wall digestibility in Arabidopsis. Biotechnol Biofuels. BioMed Central; 2015;8: 224. doi:10.1186/s13068-015-0403-z

133. Ma F, Jazmin LJ, Young JD, Allen DK. Isotopically nonstationary 13C flux analysis of changes in Arabidopsis thaliana leaf metabolism due to high light acclimation. Proc Natl Acad Sci U S A. 2014;111: 16967–72. doi:10.1073/pnas.1319485111

134. Alonso AP, Piasecki RJ, Wang Y, LaClair RW, Shachar-Hill Y. Quantifying the labeling and the levels of plant cell wall precursors using ion chromatography tandem mass spectrometry. Plant Physiol. 2010;153: 915–24. doi:10.1104/pp.110.155713

135. Matsuda F, Morino K, Miyashita M, Miyagawa H. Metabolic flux analysis of the phenylpropanoid pathway in wound-healing potato tuber tissue using stable isotope-labeled tracer and LC-MS spectroscopy. Plant Cell Physiol. 2003;44: 510–7. Available: http://www.ncbi.nlm.nih.gov/pubmed/12773637

136. Matsuda F, Morino K, Ano R, Kuzawa M, Wakasa K, Miyagawa H. Metabolic flux analysis of the phenylpropanoid pathway in elicitor-treated potato tuber tissue. Plant Cell Physiol. 2005;46: 454–466. doi:10.1093/pcp/pci042

137. Boatright J, Negre F, Chen X, Kish CM, Wood B, Peel G, et al. Understanding in Vivo Benzenoid Metabolism in Petunia Petal Tissue 1. 2011;135: 1993–2011. doi:10.1104/pp.104.045468.several

138. Heinzle E, Matsuda F, Miyagawa H, Wakasa K, Nishioka T. Estimation of metabolic fluxes, expression levels and metabolite dynamics of a secondary metabolic pathway in potato using label pulse-feeding experiments combined with kinetic network modelling and simulation. Plant J. 2007;50: 176–87. doi:10.1111/j.1365-313X.2007.03037.x

139. Rohde A, Morreel K, Ralph J, Goeminne G, Hostyn V, De Rycke R, et al. Molecular phenotyping of the pal1 and pal2 mutants of Arabidopsis thaliana reveals far-reaching consequences on phenylpropanoid, amino acid, and carbohydrate metabolism. Plant Cell. 2004;16: 2749–71. doi:10.1105/tpc.104.023705

140. Fukushima RS, Hatfield RD. Comparison of the acetyl bromide spectrophotometric method with other analytical lignin methods for determining lignin concentration in forage samples. J Agric Food Chem. 2004;52: 3713–3720. doi:10.1021/jf035497l

141. Moreira-Vilar FC, Siqueira-Soares RDC, Finger-Teixeira A, De Oliveira DM, Ferro AP, Da Rocha GJ, et al. The acetyl bromide method is faster, simpler and presents best recovery of lignin in different herbaceous tissues than klason and thioglycolic acid methods. PLoS One. 2014;9. doi:10.1371/journal.pone.0110000

142. Peng J, Lu F, Ralph J. The DFRC Method for Lignin Analysis. 4. Lignin Dimers Isolated from DFRC-Degraded Loblolly Pine Wood. J Agric Food Chem. 1998;46: 553–560. doi:10.1021/jf970802m

143. Lu F, Ralph J. Derivatization Followed by Reductive Cleavage (DFRC Method), a New Method for Lignin Analysis: Protocol for Analysis of DFRC Monomers. J Agric Food Chem. 1997;45: 2590–2592. doi:10.1021/jf970258h

144. Lu F, Ralph J. DFRC Method for Lignin Analysis. 1. New Method for -Aryl Ether Cleavage: Lignin Model Studies. J Agric Food Chem. 1997;45: 4655–4660. doi:10.1021/jf970539p

145. Ehlting J, Büttner D, Wang Q, Douglas CJ, Somssich IE, Kombrink E. Three 4-coumarate:coenzyme A ligases in arabidopsis thaliana represent two evolutionarily divergent classes in angiosperms. Plant J. 1999;19: 9–20. doi:10.1046/j.1365-313X.1999.00491.x

146. Vanholme R, Storme V, Vanholme B, Sundin L, Christensen JH, Goeminne G, et al. A systems biology view of responses to lignin biosynthesis perturbations in Arabidopsis. Plant Cell. 2012;24: 3506–29. doi:10.1105/tpc.112.102574

147. Widhalm JR, Jaini R, Morgan JA, Dudareva N. Rethinking how volatiles are released from plant cells. Trends Plant Sci. Elsevier Ltd; 2015;20: 545–550. doi:10.1016/j.tplants.2015.06.009

148. Moinuddin SG a, Jourdes M, Laskar DD, Ki C, Cardenas CL, Kim K-W, et al. Insights into lignin primary structure and deconstruction from Arabidopsis thaliana COMT (caffeic acid O-methyl transferase) mutant Atomt1. Org Biomol Chem. 2010;8: 3928–3946. doi:10.1039/c004817h

149. Nair RB, Xia Q, Kartha CJ, Kurylo E, Hirji RN, Datla R, et al. Arabidopsis CYP98A3 Mediating Aromatic 3- Hydroxylation . Developmental Regulation of the Gene , and Expression in Yeast 1. Plant Physiol. 2002;130: 210–220. doi:10.1104/pp.008649.210

150. Chen H-C, Li Q, Shuford CM, Liu J, Muddiman DC, Sederoff RR, et al. Membrane protein complexes catalyze both 4- and 3-hydroxylation of cinnamic acid derivatives in monolignol biosynthesis. Proc Natl Acad Sci U S A. 2011;108: 21253–8. doi:10.1073/pnas.1116416109

151. Baxter CJ, Liu JL, Fernie AR, Sweetlove LJ. Determination of metabolic fluxes in a non-steady-state system. Phytochemistry. 2007;68: 2313–2319. doi:10.1016/j.phytochem.2007.04.026

152. Ahn WS, Antoniewicz MR. Metabolic flux analysis of CHO cells at growth and non-growth phases using isotopic tracers and mass spectrometry. Metab Eng. Elsevier; 2011;13: 598–609. doi:10.1016/j.ymben.2011.07.002

153. Antoniewicz MR. Dynamic metabolic flux analysis-tools for probing transient states of metabolic networks. Curr Opin Biotechnol. Elsevier Ltd; 2013;24: 973–978. doi:10.1016/j.copbio.2013.03.018

154. Martínez VS, Buchsteiner M, Gray P, Nielsen LK, Quek L-E. Dynamic metabolic flux analysis using B-splines to study the effects of temperature shift on CHO cell metabolism. Metab Eng Commun. Elsevier; 2015;2: 46–57. doi:10.1016/j.meteno.2015.06.001

155. Leighty RW, Antoniewicz MR. Dynamic metabolic flux analysis (DMFA): A framework for determining fluxes at metabolic non-steady state. Metab Eng. Elsevier; 2011;13: 745–755. doi:10.1016/j.ymben.2011.09.010

156. Bonawitz ND, Kim JI, Tobimatsu Y, Ciesielski PN, Anderson NA, Ximenes E, et al. Disruption of Mediator rescues the stunted growth of a lignin-deficient Arabidopsis mutant. Nature. Nature Publishing Group; 2014;509: 376–380. doi:10.1038/nature13084

157. Humphreys JM, Chapple C. Rewriting the lignin roadmap. Curr Opin Plant Biol. 2002;5: 224–229. doi:10.1016/S1369-5266(02)00257-1

158. Boerjan W, Ralph J, Baucher M. Lignin Biosynthesis. Annu Rev Plant Biol. 2003;54: 519–46. doi:10.1146/annurev.arplant.54.031902.134938

159. Grima-Pettenati J, Goffner D. Lignin genetic engineering revisited. Plant Sci. 1999;145: 51–65. doi:10.1016/S0168-9452(99)00051-5

160. Wang H, Xue Y, Chen Y, Li R, Wei J. Lignin modification improves the biofuel production potential in transgenic Populus tomentosa. Ind Crops Prod. Elsevier B.V.; 2012;37: 170–177. doi:10.1016/j.indcrop.2011.12.014

161. Vanholme R, Storme V, Vanholme B, Sundin L, Christensen JH, Goeminne G, et al. A systems biology view of responses to lignin biosynthesis perturbations in Arabidopsis. Plant Cell. 2012;24: 3506–29. doi:10.1105/tpc.112.102574

162. Niklas J, Schr??der E, Sandig V, Noll T, Heinzle E. Quantitative characterization of metabolism and metabolic shifts during growth of the new human cell line AGE1.HN using time resolved metabolic flux analysis. Bioprocess Biosyst Eng. 2011;34: 533–545. doi:10.1007/s00449-010-0502-y

163. Colón AM, Sengupta N, Rhodes D, Dudareva N, Morgan J. A kinetic model describes metabolic response to perturbations and distribution of flux control in the benzenoid network of Petunia hybrida. Plant J. 2010;62: 64–76. doi:10.1111/j.1365-313X.2010.04127.x

164. Lee Y, Voit EO. Mathematical modeling of monolignol biosynthesis in Populus xylem. Math Biosci. Elsevier Inc.; 2010;228: 78–89. doi:10.1016/j.mbs.2010.08.009

165. Faraji M, Fonseca LL, Escamilla-Treviño L, Dixon RA, Voit EO. Computational inference of the structure and regulation of the lignin pathway in Panicum virgatum. Biotechnol Biofuels. BioMed Central; 2015;8: 151. doi:10.1186/s13068-015-0334-8

166. Schauer N, Fernie AR. Plant metabolomics: towards biological function and mechanism. Trends Plant Sci. 2006;11: 508–516. doi:10.1016/j.tplants.2006.08.007

167. Sumner LW, Mendes P, Dixon RA. Plant metabolomics: Large-scale phytochemistry in the functional genomics era. Phytochemistry. 2003;62: 817–836. doi:10.1016/S0031-9422(02)00708-2

168. Roessner U. Metabolic Profiling Allows Comprehensive Phenotyping of Genetically or Environmentally Modified Plant Systems. Plant Cell Online. 2001;13: 11–29. doi:10.1105/tpc.13.1.11

169. Sweetlove LJ, Obata T, Fernie AR. Systems analysis of metabolic phenotypes: What have we learnt? Trends Plant Sci. Elsevier Ltd; 2014;19: 222–230. doi:10.1016/j.tplants.2013.09.005

170. Allen DK. Quantifying plant phenotypes with isotopic labeling & metabolic flux analysis. Curr Opin Biotechnol. Elsevier Ltd; 2016;37: 45–52. doi:10.1016/j.copbio.2015.10.002

171. Vogt T. Phenylpropanoid biosynthesis. Mol Plant. 2010;3: 2–20. doi:10.1093/mp/ssp106

172. Li X, Bonawitz ND, Weng J-K, Chapple C. The growth reduction associated with repressed lignin biosynthesis in Arabidopsis thaliana is independent of flavonoids. Plant Cell. 2010;22: 1620–32. doi:10.1105/tpc.110.074161

173. Tiessen A, Nerlich A, Faix B, Hümmer C, Fox S, Trafford K, et al. Subcellular analysis of starch metabolism in developing barley seeds using a non-aqueous fractionation method. J Exp Bot. 2012;63: 2071–87. doi:10.1093/jxb/err408

174. Arrivault S, Guenther M, Florian A, Encke B, Feil R, Vosloh D, et al. Dissecting the Subcellular Compartmentation of Proteins and Metabolites in Arabidopsis Leaves Using Non-aqueous Fractionation. Mol Cell Proteomics. 2014;13: 2246–2259. doi:10.1074/mcp.M114.038190

175. Krueger S, Giavalisco P, Krall L, Steinhauser M-C, Büssis D, Usadel B, et al. A topological map of the compartmentalized Arabidopsis thaliana leaf metabolome. PLoS One. 2011;6: e17806. doi:10.1371/journal.pone.0017806

176. Yamada K, Norikoshi R, Suzuki K, Imanishi H, Ichimura K. Determination of subcellular concentrations of soluble carbohydrates in rose petals during opening by nonaqueous fractionation method combined with infiltration-centrifugation method. Planta. 2009;230: 1115–27. doi:10.1007/s00425-009-1011-6

177. Farre EM, Tiessen A, Roessner U, Geigenberger P, Trethewey RN. Analysis of the Compartmentation of Glycolytic Intermediates , Nucleotides , Sugars , Organic Acids , Amino Acids , and Sugar Alcohols in Potato Tubers Using a Nonaqueous Fractionation Method 1. 2001;127: 685–700. doi:10.1104/pp.010280.1

178. Gerhardt R, Heldt HW. Freeze-Stopped Material. Plant Physiol. 1984;75: 542–547.

179. Riens B, Lohaus G, Heineke D, Heldt HW. Amino Acid and Sucrose Content Determined in the Cytosolic , Chloroplastic , and Vacuolar Compartments and in the Phloem Sap of Spinach Leaves1 Federal Republic of Germany. Plant Physiol. 1991;97: 227–233.

180. Geigenberger P, Tiessen A, Meurer J. Chloroplast Research in Arabidopsis. Jarvis RP, editor. Totowa, NJ: Humana Press; 2011;775. doi:10.1007/978-1-61779-237-3

181. Tohge T, Ramos MS, Nunes-Nesi A, Mutwil M, Giavalisco P, Steinhauser D, et al. Toward the storage metabolome: profiling the barley vacuole. Plant Physiol. 2011;157: 1469–82. doi:10.1104/pp.111.185710

182. Klie S. Analysis of the compartmentalized metabolome – a validation of the non-aqueous fractionation technique. Front Plant Sci. 2011;2. doi:10.3389/fpls.2011.00055

183. Riens B, Lohaus G, Heineke D, Heldt HW. Amino Acid and sucrose content determined in the cytosolic, chloroplastic, and vacuolar compartments and in the Phloem sap of spinach leaves. Plant Physiol. 1991;97: 227–33. doi:10.1104/pp.97.1.227

184. Naik SN, Goud V V., Rout PK, Dalai AK. Production of first and second generation biofuels: A comprehensive review. Renew Sustain Energy Rev. 2010;14: 578–597. doi:10.1016/j.rser.2009.10.003

185. Vargas L, Cesarino I, Vanholme R, Voorend W, de Lyra Soriano Saleme M, Morreel K, et al. Improving total saccharification yield of Arabidopsis plants by vessel-specific complementation of caffeoyl shikimate esterase (cse) mutants. Biotechnol Biofuels. BioMed Central; 2016;9: 139. doi:10.1186/s13068-016-0551-9

186. Zhu L, O'Dwyer JP, Chang VS, Granda CB, Holtzapple MT. Multiple linear regression model for predicting biomass digestibility from structural features. Bioresour Technol. Elsevier Ltd; 2010;101: 4971–4979. doi:10.1016/j.biortech.2009.11.034

187. Healey AL, Lee DJ, Lupoi JS, Papa G, Guenther JM, Corno L, et al. Evaluation of Relationships between Growth Rate, Tree Size, Lignocellulose Composition, and Enzymatic Saccharification in Interspecific Corymbia Hybrids and Parental Taxa. Front Plant Sci. 2016;7: 1–14. doi:10.3389/fpls.2016.01705

188. Liu B, G??mez LD, Hua C, Sun L, Ali I, Huang L, et al. Linkage mapping of stem saccharification digestibility in rice. PLoS One. 2016;11: 1–13. doi:10.1371/journal.pone.0159117

189. Smola AJ, Sch??lkopf B. A tutorial on support vector regression. Stat Comput. 2004;14: 199–222. doi:10.1023/B:STCO.0000035301.49549.88

190. Cortes C, Vapnik V. Support-Vector Networks. Mach Learn. 1995;20: 273–297. doi:10.1023/A:1022627411411

191. Ben-Hur A, Ong CS, Sonnenburg S, Sch??lkopf B, R??tsch G. Support vector machines and kernels for computational biology. PLoS Comput Biol. 2008;4. doi:10.1371/journal.pcbi.1000173

192. Shi J, Pattathil S, Parthasarathi R, Anderson NA, Kim JI, Venketachalam S, et al. Impact of engineered lignin composition on biomass recalcitrance and ionic liquid pretreatment efficiency. Green Chem. Royal Society of Chemistry; 2016;18: 4884–4895. doi:10.1039/C6GC01193D

193. Berthet S, Demont-Caulet N, Pollet B, Bidzinski P, Cezard L, Le Bris P, et al. Disruption of LACCASE4 and 17 Results in Tissue-Specific Alterations to Lignification of Arabidopsis thaliana Stems. Plant Cell. 2011;23: 1124–1137. doi:10.1105/tpc.110.082792

194. Eudes A, Pereira JH, Yogiswara S, Wang G, Teixeira Benites V, Baidoo EEK, et al. Exploiting the substrate promiscuity of Hydroxycinnamoyl-CoA:Shikimate Hydroxycinnamoyl Transferase to reduce lignin. Plant Cell Physiol. 2016;57: 568–579. doi:10.1093/pcp/pcw016

195. Drucker H, Burges CJC, Kaufman L, Smola A, Vapnik V. Support vector regression machines. Adv Neural Inf Process Syst. 1997;1: 155–161. doi:10.1.1.10.4845

196. Signal I, Magazine P. What are Genetic Algorithms ? Optimization Algorithms. IEEE Signal Process Mag. 1996; 22–37. doi:10.1109/79.543973

197. Zang H, Zhang S, Hapeshi K. A review of nature-inspired algorithms. J Bionic Eng. Jilin University; 2010;7: S232–S237. doi:10.1016/S1672-6529(09)60240-7

198. Burke EK, Graham K. Search methodologies: Introductory tutorials in optimization and decision support techniques, second edition. Search Methodol Introd Tutorials Optim Decis Support Tech Second Ed. 2014; 1–716. doi:10.1007/978-1-4614-6940-7

199. Campbell MM, Sederoff RR. Variation in Lignin Content and Composition. Plant Physiol. 1996;1996: 3–13. doi:10.1104/pp.110.1.3

200. Efron B. Better Bootstrap Confidence Intervals. J Am Stat Assoc. 2014;82: 171–185.

201. Efron B, Tibshirani R. [Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy]: Rejoinder. Stat Sci. 1986;1: 77–77. doi:10.1214/ss/1177013817

202. Sluiter J, Sluiter A. Summative Mass Closure: Laboratory Analytical Procedure (LAP) Review and Integration: Feedstocks; Issue Date: April 2010; Revision Date: July 2011 (Version 07-08-2011). Nrel L. 2010;2011: 1–10.

203. Breu F, Guggenbichler S, Wollmann J. Machine Learning ECML 2004 [Internet]. 15th European Conference on Machine Learning Pisa, Italy, September 2004 Proceedings. 2004. doi:10.1007/b100702

204. Chawla N V., Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res. 2002;16: 321–357. doi:10.1613/jair.953

205. Japkowicz N. The Class Imbalance Problem: Significance and Strategies. Proc 2000 Int Conf Artif Intell. 2000; 111--117. doi:10.1.1.35.1693

206. Wang BX, Japkowicz N. Boosting support vector machines for imbalanced data sets. Knowl Inf Syst. 2010;25: 1–20. doi:10.1007/s10115-009-0198-y

207. Czarnecki WM, Podlewska S, Bojarski AJ. Robust optimization of SVM hyperparameters in the classification of bioactive compounds. J Cheminform. Springer International Publishing; 2015;7: 1–15. doi:10.1186/s13321-015-0088-0

208. Franke R, Hemm MR, Denault JW, Ruegger MO, Humphreys JM, Chapple C. Changes in secondary metabolism and deposition of an unusual lignin in the ref8 mutant of Arabidopsis. Plant J. 2002;30: 47–59. doi:10.1046/j.1365-313X.2002.01267.x

209. Pompon D, Louerat B, Bronine A, Urban P. Yeast expression of animal and plant P450s in optimized redox environments. Methods Enzymol. 1996;272: 51–64. doi:10.1016/S0076-6879(96)72008-6

210. Krishnakumar V, Hanlon MR, Contrino S, Ferlanti ES, Karamycheva S, Kim M, et al. Araport: The Arabidopsis Information Portal. Nucleic Acids Res. 2015;43: D1003–D1009. doi:10.1093/nar/gku1200

211. Price ND. Genome-scale modeling for metabolic engineering Evangelos. J Ind Microb Biotechnol\. 2015;42: 327–338. doi:10.1007/s10295-014-1576-3.Genome-scale

212. Zomorrodi AR, Suthers PF, Ranganathan S, Maranas CD. Mathematical optimization applications in metabolic networks. Metab Eng. Elsevier; 2012;14: 672–686. doi:10.1016/j.ymben.2012.09.005

213. Patil KR, Rocha I, Forster J, Nielsen J. Evolutionary programming as a platform for in silico metabolic engineering. BMC Bioinformatics. 2005;6: 308. doi:10.1186/1471-2105-6-308

214. Rocha I, Maia P, Rocha M, Ferreira EC. OptGene – a framework for in silico metabolic engineering. Proc Natl Acad Sci U S A. 2008; 218–219.

215. Pharkya P, Maranas CD. An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems. Metab Eng. 2006;8: 1–13. doi:10.1016/j.ymben.2005.08.003

216. Yang L, Cluett WR, Mahadevan R. EMILiO: A fast algorithm for genome-scale strain design. Metab Eng. Elsevier; 2011;13: 272–281. doi:10.1016/j.ymben.2011.03.002

217. de Oliveira Dal'Molin CG, Quek L-E, Palfreyman RW, Brumbley SM, Nielsen LK. AraGEM, a Genome-Scale Reconstruction of the Primary Metabolic Network in Arabidopsis. Plant Physiol. 2010;152: 579–589. doi:10.1104/pp.109.148817