Max-Planck-Institut für Molekulare Pflanzenphysiologie
und Universität Potsdam

Context-specific metabolic predictions: computational methods and applications

Dissertation

zur Erlangung des akademischen Grades *doctor rerum naturalium*
(Dr. rer. nat.) in der Wissenschaftsdisziplin Systembiologie

eingereicht in kumulativer Form an der Mathematisch-
Naturwissenschaftlichen Fakultät der Universität Potsdam

von

Semidán Robaina Estévez

Potsdam den 21 März 2017

Published online at the
Institutional Repository of the University of Potsdam:
URN urn:nbn:de:kobv:517-opus4-401365
http://nbn-resolving.de/urn:nbn:de:kobv:517-opus4-401365

Dedicada a mi familia.

“Keep moving forward!”

Cornelius Robinson

Contents

Acknowledgements i
2 Abstract ii
3 Zusammenfassung iii
4 List of abbreviations iv
1 Introduction 1
1.1 Metabolism 2
1.2 Models of metabolism in the pre-omics era 3
1.2.1 The mass action formalism 3
1.2.2 The Michaelis-Menten formalism 4
1.2.3 A generalization of the mass action formalism 6
1.3 Models of metabolism in the omics era 6
1.3.1 Genome-scale metabolic model reconstructions 8
1.3.2 Genome-scale metabolic model predictions 10
1.3.3 The rationale of constraint-based approaches 11
1.3.4 Brief notes on convex optimization 15
1.4 Context-specific metabolic predictions 20
1.4.1 Main types of experimental data integrated in genome-scale models 22
1.4.2 Mapping data into genome-scale models 24
1.4.3 A classification of existing methods 26
1.4.4 Alternative optimal solutions in convex optimization 29
1.4.5 Alternative optima in context-specific metabolic predictions 30
1.5 Thesis outline 31
2 Context-specific metabolic predictions with RegrEx 33
2.1 Introduction 35
2.2 Results & Discussion 37
2.2.1 The RegrEx method 37
2.2.2 Evaluation of RegrEx performance 39
2.2.3 Main characteristics of extracted models by the evaluated methods 41
2.2.4 Similarity to data evaluation 43

2.2.5 Evaluation of the models with human protein profiles 46
2.2.6 Functional analysis of RegrEx extracted models 47
2.2.7 Computation time comparison 52
2.3 Conclusion 54
2.4 Methods 55
2.4.1 RegrEx implementation.....55
2.4.2 Context-specific model extraction.....56
2.4.3 Performance analysis with competing methods 56
2.4.4 Model agreement with human protein expression data.....57
3 Alternative optima in context-specific metabolic predictions..... 58
3.1 Introduction..... 60
3.2 Results and discussion 61
3.2.1 Evaluation of alternative optima: Computational methods 61
3.2.2 Evaluation of alternative optima: Case studies 70
3.3 Conclusions..... 81
3.4 Methods 82
3.4.1 $RegrEx_{LAD}$, $RegrEx_{AOS}$, CorEx and AltNet implementations 82
3.4.2 Metabolic model and gene expression data.....82
3.4.3 Extreme flux values of the flux cone.....83
3.4.4 Sampling flux distributions from the flux cone.....84
3.4.5 Quantification of the $RegrEx_{LAD}$ alternative optima space 84
3.4.6 Measuring the distance between alternative optimal networks.....85
3.4.7 Generation of a ranked list of metabolic pathways 85
3.4.8 Functional testing of the liver-specific reconstructions.....85
4 Applying RegrEx to investigate guard cell metabolism..... 87
4.1 Introduction..... 89
4.2 Results and discussion 91
4.2.1 Computational workflow and rationale for model-driven predictions of differences in G and M cell metabolism..... 91
4.2.2 Interplay between the tricarboxylic acid cycle (TCA) cycle and PEPc in the synthesis of cytosolic malate 92
4.2.3 Chloroplasts adapt their function to meet the metabolic requirements of G cells……..94
4.2.4 The CBC drives sucrose and starch syntheses in G cells 96

4.2.5 Robustness of prediction to adding constraints derived from experimental observations 98
4.2.6 Validation of model predictions 99
4.2.7 G cells have higher anaplerotic CO_2 fixation 99
4.2.8 Guard cells have higher ^{13}C-enrichment but lower capacity to produce sucrose under ^{13}C-$NaHCO^-_3$ 101
4.3 Conclusions 103
4.4 Material and methods 104
4.4.1 Gene expression data 104
4.4.2 Metabolic network model 104
4.4.3 Gene expression integration in AraCORERed 105
4.4.4 Evaluation of the alternative optima space 105
4.4.5 Evaluation of flux values across the alternative optima space 106
4.4.6 Evaluation of flux-sum values across the alternative optima space 107
4.4.7 Integration of additional constraints derived from experimental observations 107
4.4.8 Plant material and growth conditions 108
4.4.9 Experimental set-up for in vivo G and M cells analyses 108
4.4.10 ^{13}C isotope labelling experiment using isolated M cell protoplasts 109
4.4.11 ^{13}C kinetic isotope labelling experiment in G cells 110
4.4.12 Extraction and analysis of metabolites 110
5 Discussion 112
5.1 Further considerations on Chapter 2 113
5.2 Further considerations on Chapter 3 115
5.3 Further considerations on Chapter 4 118
5.4 Future directions 119
6 Supplementary Information 122
7 Bibliography 144

Acknowledgements

I wish to express my gratitude to all the people that have helped me during the development of this thesis. First, I thank my supervisor, Zoran, for receiving me in his group, for being always there when I needed to, and for all those countless moments of excellent discussions. I also thank all present and former members of the *Systems Biology and Mathematical Modeling* group: David Breuer, Sabrina Kleessen, Nadine Töpfter, Anne Arnold, Nooshin Omranian, Jeanne Marie Onana-Eloundu Mbebi, Max Sajitz-Hermstein, Jost Neigenfind, Verónika Ceballos Núñez, Alberto Castellini, Andreas Krug, Marina Leer, Anika Küken, Kevin Schwahn, Alessio Milanese, Marko Karbevski, Hao Tong, Michael Scheunemann, Georg Basler and Jacqueline Nowak, for the warm and friendly atmosphere that they brought to the group, as well as for the many interesting and fun conversations and moments that we shared.

I thank the members of my doctoral advisory committee: Alisdair Fernie, Marek Mutwil and Lothar Willmitzer for the encouraging conversations and for their commitment to the task. I thank Ina Talke for her support and help whenever I had a question related to the PhD program, and I also thank Birgit Schäfer and Jacqueline Sommer for their support whenever I needed help organizing conference trips, or even with administrative tasks the first days after arriving in Germany. I thank the IT department for taking good care of our valuable data storage and computer devices. I thank the whole Max Planck Institute of Molecular Plant Physiology, the International Max Planck Research School *Primary Metabolism and Plant Growth* and the Max Planck Society for giving me the opportunity of fulfilling these PhD studies.

I thank all researchers whose previous discoveries and contributions made the development of the studies in this thesis possible. I also want to thank all my teachers, who encouraged me to be curious and to pursue a scientific career, special thanks to my primary school teacher, Heriberta, who believed in me and transmitted to me an eager to learn, and also to my Master thesis supervisor, Néstor Torres Darias, who introduced me into the field of Systems and Computational Biology. Finally, I thank my parents, my brother, and all my family for their unconditional love and support during good and not so good times.

To all, thank you!

Abstract

All life-sustaining processes are ultimately driven by thousands of biochemical reactions occurring in the cells: the metabolism. These reactions form an intricate network which produces all required chemical compounds, *i.e.,* metabolites, from a set of input molecules. Cells regulate the activity through metabolic reactions in a context-specific way; only reactions that are required in a cellular context, *e.g.*, cell type, developmental stage or environmental condition, are usually active, while the rest remain inactive. The context-specificity of metabolism can be captured by several kinds of experimental data, such as by gene and protein expression or metabolite profiles. In addition, these context-specific data can be assimilated into computational models of metabolism, which then provide context-specific metabolic predictions.

This thesis is composed of three individual studies focussing on context-specific experimental data integration into computational models of metabolism. The first study presents an optimization-based method to obtain context-specific metabolic predictions, and offers the advantage of being fully automated, *i.e.,* free of user defined parameters. The second study explores the effects of alternative optimal solutions arising during the generation of context-specific metabolic predictions. These alternative optimal solutions are metabolic model predictions that represent equally well the integrated data, but that can markedly differ. This study proposes algorithms to analyze the space of alternative solutions, as well as some ways to cope with their impact in the predictions.

Finally, the third study investigates the metabolic specialization of the guard cells of the plant *Arabidopsis thaliana,* and compares it with that of a different cell type, the mesophyll cells. To this end, the computational methods developed in this thesis are applied to obtain metabolic predictions specific to guard cell and mesophyll cells. These cell-specific predictions are then compared to explore the differences in metabolic activity between the two cell types. In addition, the effects of alternative optima are taken into consideration when comparing the two cell types. The computational results indicate a major reorganization of the primary metabolism in guard cells. These results are supported by an independent ^{13}C labelling experiment.

Zusammenfassung

Alle lebenserhaltenden Prozesse werden durch tausende biochemische Reaktionen in der Zelle bestimmt, welche den Metabolismus charakterisieren. Diese Reaktionen bilden ein komplexes Netzwerk, welches alle notwendigen chemischen Verbindungen, die sogenannten Metabolite, aus einer bestimmten Menge an Ausgangsmolekülen produziert Zellen regulieren ihren Stoffwechsel kontextspezifisch, dies bedeutet, dass nur Reaktionen die in einem zellulären Kontext, zum Beispiel Zelltyp, Entwicklungsstadium oder verschiedenen Umwelteinflüssen, benötigt werden auch tatsächlich aktiv sind. Die übrigen Reaktionen werden als inaktiv betrachtet. Die Kontextspezifität des Metabolismus kann durch verschiedene experimentelle Daten, wie Gen- und Proteinexpressionen oder Metabolitprofile erfasst werden. Zusätzlich können diese Daten in Computersimulationen des Metabolismus integriert werden, um kontextspezifische (metabolische) Vorhersagen zu treffen.

Diese Doktorarbeit besteht aus drei unabhängigen Studien, welche die Integration von kontextspezifischen experimentellen Daten in Computersimulationen des Metabolismus thematisieren. Die erste Studie beschreibt ein Konzept, basierend auf einem mathematischen Optimierungsproblem, welches es erlaubt kontextspezifische, metabolische Vorhersagen zu treffen. Dabei bietet diese vollautomatische Methode den Vorteil vom Nutzer unabhängige Parameter, zu verwenden. Die zweite Studie untersucht den Einfluss von alternativen optimalen Lösungen, welche bei kontextspezifischen metabolischen Vorhersagen generiert werden. Diese alternativen Lösungen stellen metabolische Modellvorhersagen da, welche die integrierten Daten gleichgut wiederspiegeln, sich aber grundlegend voneinander unterscheiden können. Diese Studie zeigt verschiedene Ansätze alternativen Lösungen zu analysieren und ihren Einfluss auf die Vorhersagen zu berücksichtigen.

Schlussendlich, untersucht die dritte Studie die metabolische Spezialisierung der Schließzellen in *Arabidopsis thaliana* und vergleicht diese mit einer weiteren Zellart, den Mesophyllzellen. Zu diesem Zweck wurden die in dieser Doktorarbeit vorgestellten Methoden angewandt um metabolische Vorhersagen speziell für Schließzellen und Mesophyllzellen zu erhalten. Anschließend wurden die zellspezifischen Vorhersagen auf Unterschiede in der metabolischen Aktivität der Zelltypen, unter Berücksichtigung des Effekt von alternativen Optima, untersucht. Die Ergebnisse der Simulationen legen eine grundlegende Neuorganisation des Primärmetabolismus in Schließzellen verglichen mit Mesophyllzellen nahe. Diese Ergebnisse werden durch unabhängige ^{13}C markierungs Experimente bestätigt.

List of abbreviations

ABA	Abscisic acid
AMP	Adenosine monophosphate
ATP	Adenosine triphosphate
CA	Carbonic anhydrase
CBAs	Constraint-based approaches
CBC	Calvin-Benson cycle
cDNA	Complementary DNA
CoA	Coenzyme A
CTP	Cytidine triphosphate
CV	Coefficient of variation
DHAP	Dihydroxyacetone phosphate
DNA	Deoxyribonucleic acid
FBA	Flux balance analysis
FD	Ferredoxin
FD^-	Reduced ferredoxin
FPKM	Fragments per kilobase per million mapped reads
FVA	Flux variability analysis
G1P	Glucose 1-phosphate
G3P	Glyceraldehyde 3-phosphate
G6P	Glucose 6-phosphate
GCMS	Gas chromatography gas spectrometry
GEM	Genome-scale metabolic model
Gln	Glutamine
Glu	Glutamate
GMP	Guanosine monophosphate
GPR rules	Gene-protein-reaction rules
HXK	Hexokinase

Inv	Invertase
LP	Linear program
Mal	Malate
MAS5	Microarray analysis suite 5
MDH	Malate dehydrogenase
MFC	Mean flux capacity
MILP	Mixed integer linear program
MIQP	Mixed integer quadratic program
mRNA	Messenger RNA
NAD	Nicotinamide adenine dinucleotide
NADP	Nicotinamide adenine dinucleotide phosphate
OAA	Oxaloacetate
PEP	Phosphoenolpyruvate
PEPc	Phosphoenolpyruvate carboxylase
PGA	Phosphoglyceric acid
Pyr	Pyruvate
QP	Quadratic program
RMA	Robust multi-array average
RNA	Ribonucleic acid
RNA-Seq	RNA sequencing
ROS	Reactive oxygen species
RPKM	Reads per kilobase per million mapped reads
RuBisCO	Ribulose-1,5-bisphosphate carboxylase/oxygenase
SPS	Sucrose-phosphate synthase
Suc	Sucrose
SuSy	Sucrose synthase
TCA	Tricarboxylic acid cycle
TNR	Total number of reactions
UDP	Uridine diphosphate
α – KG	α - Ketoglutarate

Chapter 1

Introduction

An intricate network of biochemical reactions lies at the core of all cellular processes: the metabolism. Metabolism transforms environmental energy and matter into useful forms for the cells, and interconverts chemicals, *i.e.*, *metabolites*, to create all necessary compounds sustaining life. In addition, metabolism operates in a specialized manner, adapting the activity trough the reactions to meet the requirements of diverse cellular contexts—*e.g.*, cell type, developmental stage or environmental conditions. A thorough understanding of the adaptation of metabolism is therefore basic to deciphering how cells operate. In this respect, experimental techniques in biochemistry and molecular biology provide a wealth of context-specific data. These data can be combined with mathematical and computational descriptions of metabolism, which provide a quantitative frame to interpret experimental observations, test diverse hypotheses and guide further experiments.

This thesis provides mathematical and computational methods which integrate experimental data to obtain context-specific metabolic predictions, and it is composed of three related studies. The first two studies present the new methods, while the third study applies these methods to investigate how metabolism specializes in a concrete cell type: the guard cells of the plant *Arabidopsis thaliana.* Each study is presented in an individual chapter and provides detailed background. In contrast, this introductory chapter provides a general background to the subject of this thesis. We will begin by elaborating on the definition of metabolism, which will be followed by a brief historical description of the development of mathematical models of metabolism. We will continue with an introduction to the rationale behind the class of methods developed in this thesis. We will finalize this chapter with the main types of experimental data used to obtain context-specific metabolic predictions, followed by a summary of the state-of-the-art computational methods devoted to this end.

1.1 Metabolism

Cellular metabolism is composed of thousands of (bio)chemical reactions and metabolites. Reactions and metabolites do not occur isolated, but form an intricated network in which metabolites can be shared between several reactions. Moreover, the reactions act in an organized and controlled way to meet the cellular requirements. On the one hand, reactions are thermodynamically coupled, that is, reactions that release energy are coupled to reactions that require energy to proceed. The thermodynamic coupling enables the activity through reactions that would never occur in isolation, and renders an overall feasible metabolism (Voet & Voet, 2011). On the other hand, the activity through most reactions is coordinated by proteins or protein complexes, the enzymes, which are biological catalyzers. This is because most metabolic reactions, even though thermodynamically feasible, would proceed at very slow rates without catalysis.

Enzymes not only speed up metabolism, but also provide an entry point to regulatory processes, which control the flux[1] through reactions. This control is exerted by varying the concentration of functionally active enzymes, which can be accomplished by altering: *i*) the transcription of the coding genes and/or the stability of the corresponding transcripts, *ii*) the translation rate into proteins and/or the stability of the proteins, *iii*) the inactivation of existing enzymes, either by means of allosteric regulation, posttranslational modifications or by controlling the availability of cofactors essential to enzyme functioning (Metallo & Vander Heiden, 2013).

Although metabolism has not a clear natural partitioning, groups of metabolic reactions have been traditionally classified based on general physiological functions. For instance, metabolism can be decomposed into *catabolism* and *anabolism*: Catabolic reactions typically break high energy compounds into smaller molecules and obtain in the process free energy. In contrast, anabolic reactions typically synthetize complex compounds from simpler ones; a process that utilizes the free energy derived from catabolic reactions or directly withdrawn from the environment (Voet & Voet, 2011). A finer partition of metabolism employs the concept of *metabolic pathway*. Metabolic pathways are chains of linked reactions, in which at least one product of a reaction acts as a substrate of the next one in the chain. The reactions in a pathway generally act together to accomplish a specific function within metabolism, and may be switched on or off depending on the metabolic demands of a cellular context. In this sense, metabolism can be regarded as a collection of metabolic pathways specialized in certain metabolic tasks. An example of a catabolic pathway is the *glycolysis*, a sequence of 10 reactions that collectively break down glucose into smaller subunits; a process that releases free energy which can be employed to drive anabolic processes (Bar-Even, Flamholz, Noor, & Milo, 2012).

[1] The flux of a reaction corresponds to the rate at which a substrate is consumed (or a product synthetized), as measured, for instance, by moles of consumed substrate per volume per unit time.

1.2 Models of metabolism in the pre-omics era

Most of the current scientific knowledge of metabolism derives from studies conducted throughout the 20th century, which employed classical techniques from biochemistry. During this time, the investigations were tailored to discovering new metabolic pathways or characterizing isolated reactions (Voet & Voet, 2011). These efforts produced a wealth of information about enzyme kinetics and reaction mechanisms, which vastly increased the scientific knowledge of metabolism. However, during this time, metabolic pathways were generally studied in isolation, like small submodules of metabolism specialized in fulfilling a particular function. This approach sharply contrasts to the modern, systemic approach to studying metabolism of the "omics" era, which will be later discussed.

1.2.1 The mass action formalism

The advent of a wealth of biochemical data encouraged the development of the first mathematical models of metabolism. In 1913, Leonor Michaelis and Maud Menten proposed the first formalism explicitly developed to model enzymatically catalyzed (biochemical) reactions, the Michaelis-Menten rate law (K. A. Johnson, Goody, Johnson, & Goody, 2011). However, this formalism was derived from the mass action rate law, a previous formalism developed to model non-enzymatically catalyzed reactions. The law of mass action, originally conceived by Cato Guldberg and Peter Waage in 1864 (Voit, Martens, & Omholt, 2015), is a probabilistic law stating that the flux through a reaction is proportional to the product of the concentrations of the substrates. For instance, consider the following reaction.

$$2X_1 + X_2 \underset{v_2}{\overset{v_1}{\rightleftharpoons}} X_3 \tag{1.1}$$

where two molecules of the chemical species X_1 react with a molecule of X_2 to form X_3. To simplify notation, we will abuse language from now on and denote the concentration (*e.g.* in molar units) of a species by the same denominating symbols, then the flux, *i.e.,* reaction rate, v_1 in the forward direction of the reaction (from left to right) corresponds to

$$v_1 = k_1 X_1^2 X_2 \tag{1.2}$$

and the flux v_2 of the backward direction to

$$v_2 = k_2 X_3 \tag{1.3}$$

where the exponents are the stoichiometric coefficients of the corresponding species in the reaction, and the kinetic constants, k_1, k_2, are parameters that depend on the

particular reaction mechanism and environmental conditions. Therefore, for a general reaction, the flux, *v,* under the law of mass action is represented by

$$v = k\prod_{i=1}^{m} X_i^{|s_{ij}|} \tag{1.4}$$

where s_i is the stoichiometric coefficient of each reactant species X_i in the reaction.

1.2.2 The Michaelis-Menten formalism

Leonor Michaelis and Maud Menten began considering the simple, enzymatically catalyzed reaction

$$E + X_1 \underset{v_2}{\overset{v_1}{\rightleftharpoons}} Z \xrightarrow{v_3} E + X_2 \tag{1.5}$$

which is a two-step reaction. The first step involves the binding of one molecule of substrate X_1 to a molecule of free enzyme E to form the enzyme-substrate complex Z, a process that is reversible. In the second step, the product X_2 is formed and released from the enzyme, a process that is assumed irreversible. To see how the Michaelis-Menten formalism derives from the law of mass action, we begin by considering the dynamical system

$$\begin{aligned} \dot{E} &= -k_1 E X_1 + (k_2 + k_3) Z \\ \dot{X}_1 &= -k_1 E X_1 + k_2 Z \\ \dot{Z} &= k_1 E X_1 - (k_2 + k_3) Z \\ \dot{X}_2 &= k_3 Z \end{aligned} \tag{1.6}$$

which is represented under the mass action formalism—first order time derivatives are symbolized by a dot over the variable name. Michaelis-Menten further assumes that the first step of the reaction occurs much more rapidly than the second. Originally, the first step was assumed to be at equilibrium, so that $k_1 E X_1 = k_2 Z$ hence $\dot{X}_1 = 0$ (Voit et al., 2015). Alternatively, in 1925, G.E. Briggs and J.B.S. Haldane (Briggs & Haldane, 1925) proposed the quasi-steady-state assumption in which

$$\dot{Z} = k_1 E \dot{X}_1 - (k_2 + k_3) Z = 0 \tag{1.7}$$

so that

$$k_1 E \dot{X}_1 = (k_2 + k_3) Z \tag{1.8}$$

and which we will follow in the derivation. We also note that the system (1.6) has the conserved quantity

$$E + Z = E_T \tag{1.9}$$

since we have $\dot{E} + \dot{Z} = 0$. We continue the derivation by realizing that the global flux, converting X_1 into X_2, in (1.5) effectively depends on $v_3 = k_3 Z$, since, by assumption, it is the slowest step. However, the enzyme-substrate complex is unstable and hence unmeasurable experimentally, thus our final goal is to express Z as a function of measurable variables. By (1.8) we obtain

$$\frac{EX_1}{Z} = \frac{k_2 + k_3}{k_1} \tag{1.10}$$

and from (1.8) and (1.9) we have

$$\frac{(E_T - Z)X_1}{Z} = \frac{k_2 + k_3}{k_1} \ . \tag{1.11}$$

Hence

$$Z = \frac{E_T X_1}{K_M + X_1} \tag{1.12}$$

with

$$K_M = \frac{k_2 + k_3}{k_1} . \tag{1.13}$$

Finally, we arrive at the desired expression

$$v_3 = \frac{v_{\max} X_1}{K_M + X_1} , \tag{1.14}$$

which expresses the Michaelis-Menten rate function for the reaction in (1.5). The expression in (1.14) depends on the concentration of the substrate S and on two parameters: $v_{\max} = k_3 E_T$ corresponds to the maximum rate of the reaction, and depends on the catalytic constant k_3, usually denoted k_{cat}, the maximum turnover of the enzyme—*i.e.,* maximum conversion rate of substrate into product—and the total concentration of enzyme, E_T, in the system. On the other hand K_M represents an inverse measure of the affinity of the enzyme for its substrate, and can be defined as the concentration of the substrate in the system when $v_3 = ½ v_{\max}$ (Voet & Voet, 2011). Both parameters can be measured *in vitro* and serve to characterize the catalytic properties of the enzyme in the reaction under study.

Contrary to the mass action formalism, the Michaelis-Menten rate law does not have a general form; its derivation depends on the type of biochemical reaction being modelled and on biochemical assumptions, such as the quasi steady-state assumption on the enzyme-substrate complex. In fact, the simple formulation in (1.14) becomes more convoluted with more complex reactions. For instance, in reactions with more

than one substrate, the different orders in which each substrate is utilized render different formulations of the rate law (Leskovac, 2003). Additionally, the affinity for a substrate, captured by K_M, may not be constant, but a function of the same or other substrates in the system. The Hill equation (Goutelle et al., 2008) extends Michaelis-Menten to account for cases in which the binding of a molecule of substrate increases of decreases the affinity of the enzyme, a process known as positive and negative cooperative binding, respectively. On the other hand, other extensions of (1.14) are necessary when non-substrate molecules compete for the active site of the enzyme, thus inhibiting the rate of the reaction and affecting v_{max}, which is no longer constant (Leskovac, 2003).

1.2.3 A generalization of the mass action formalism

The Michaelis-Menten formalism and its extensions provide more realistic predictions than that of the law of mass action. However, this comes at the cost of losing a canonical representation of the formalism and of more free parameters to determine, which restricts the usage of Michaelis-Menten to well-studied and small-scale biochemical systems. Motivated by the need to address this problem, generalizations of the mass action law were proposed in the 1970s. For instance, one of this formalisms, is the *generalized mass action* formalism (Savageau, 1969), in which the reaction rate for a general reaction is defined as

$$v = \gamma \prod_{i=1}^{m} X_i^{g_i} \ . \tag{1.15}$$

The only differences between (1.15) and (1.4) are (*i*) the exponents g_i are not the stoichiometric coefficients, but parameters of the system which can take any real value within an allowable range, (*ii*) other participating species, besides reactants, may be included, such as inhibitory species, in which case the corresponding exponent $g_i < 0$. These characteristics make the generalized mass action formalism suitable to model phenomenological processes instead of single biochemical reactions. For instance, v may represent the net flux through an entire metabolic pathway, in which several reactions are grouped together into a single process. Therefore, this formalism may be interpreted as a compromise between the simple but limited law of mass action, and the biochemically more realistic but also complex Michaelis-Menten.

1.3 Models of metabolism in the omics era

The development of faster and automated sequencing techniques led to a milestone in molecular biology: the sequencing of the first (non-viral) genome, that of the bacterial species *Haemophilus influenza* in 1995 (Fleischmann et al., 1995). Subsequent improvements in sequencing techniques rapidly increased the number of species with a fully-sequenced genome. Furthermore, the development of bioinformatics tools

allowed annotating extensive lists of genes, as well as unravelling putative gene functions based on sequence comparison with known genes (Stein, 2001). Altogether, these technological innovations enabled large-scale analysis of whole genomes, thus starting the *genomics* era. Thousands of genomes have been sequenced since 1995—including the human genome in 2003 (Human Genome Sequencing Consortium, 2004)—and are available in several online databases, which provide easy and public access to this wealth of information, *e.g.* (Auton et al., 2015; Berardini et al., 2015; Caspi et al., 2016; Kanehisa, Sato, Kawashima, Furumichi, & Tanabe, 2016; Yates et al., 2016).

The paradigm of large-scale or systemic analyses of organisms did not stop with genomics, but continued to be applied to other areas of molecular biology. New techniques have allowed the quantification of the expression of thousands of annotated genes in a single experiment (Butte, 2002; Z. Wang, Gerstein, & Snyder, 2009). This *transcriptomics* data, give a global picture of the gene expression state, *i.e.,* the average number of transcripts for the monitored genes in a cell. Therefore, while genomics provides a static vision of the genome, transcriptomics captures how the genome is being effectuated by the cell at the time of the experiment, and hints at new possible physiological functions.

Today, modern high-throughput experimental techniques allow taking large-scale measurements of additional elements in the hierarchical causal processes controlling metabolism—*i.e.,* from gene transcription to messenger RNA (mRNA) translation into proteins (enzymes), to posttranslational modification, and to the reaction fluxes and the metabolites. Proteomics characterize and quantify thousands of proteins at the same time (Righetti, Campostrini, Pascali, Hamdan, & Astner, 2004). Similarly, metabolomics aims at identifying and measuring the concentrations of a large number of metabolites (C. H. Johnson, Ivanisevic, & Siuzdak, 2016; Rochfort, 2005). More recently, attempts have been made into estimating the metabolic fluxes, although current techniques only allow the quantification of a relatively small number of metabolic fluxes in central metabolism (Niedenführ, Wiechert, & Nöh, 2015).

The investigation of metabolism changed drastically upon the arrival of the omics data: from the classical small-scale and pathway-centered approach to the modern systemic and network-centered. This change has required the development of new computational and statistical techniques to preprocess, analyze and integrate the different kinds of omics data. In this sense, genome-scale metabolic models and constraint-based approaches have proven successful computational methods to investigate large-scale properties of metabolism, both at a basic level and in applied research (Cook & Nielsen, 2017; Oberhardt, Palsson, & Papin, 2009; Zhang & Hua, 2015)

1.3.1 Genome-scale metabolic model reconstructions

Genome-scale metabolic models (GEMs) are data structures that aim at integrating the metabolic network of a species with additional biochemical information (Kim, Sohn, Kim, & Kim, 2012). GEMs list all known biochemical reactions and provide a detailed stoichiometry of the consumed and produced metabolites, which fully describes the underlying metabolic network. Furthermore, additional data associated to each reaction are included. Reactions that are enzymatically catalyzed are associated to an enzyme commission (EC) number (International Union of Biochemistry and Molecular Biology. Nomenclature Committee. & Webb, 1992), which classifies biochemical reactions based on their mechanism of action. In most cases, the conventional protein and gene identifiers associated to each reaction are also provided, as well as the thermodynamic reversibility of each reaction under normal physiological conditions.

Most GEMs of eukaryotic species are also compartmentalized. That is, reactions in the model are grouped into the major subcellular compartments like: the cytosol, mitochondrion, vacuole, peroxisome and the chloroplast (if applicable). Each compartment then operates as a metabolic subnetwork which is connected to the rest of the compartments by suitable transport reactions. Altogether, these features make GEMs excellent sources of metabolic information. However, the principal utility of GEMs—as the name suggests—is to serve as models of metabolism, *i.e.*, provide *in silico* metabolic predictions under particular scenarios.

The first GEM, that of the bacterium *Haemophilus influenzae*, appeared in 1999 result of the pioneering work of Benhard Palsson's group (Edwards & Palsson, 1999). This model was rapidly followed by other GEMs of bacterial species, such as *Escherichia coli* in the year 2000 and *Helicobacter pylori* in 2002. In 2003 the first GEM of a eukaryotic organism, the yeast *Saccharomyces cerevisiae*, was released (Förster, Famili, Fu, Palsson, & Nielsen, 2003), to which it followed the first GEMs of complex multicellular organisms: mouse in 2005 (Sheikh, Förster, & Nielsen, 2008) and human in 2007 (Duarte et al., 2007). The first GEM of a plant species, namely *Arabidopsis thaliana*, appeared in 2009 (Poolman, Miguet, Sweetlove, & Fell, 2009).

Today, hundreds of metabolic models of species from the three domains of life are available and accessible through public databases (King et al., 2016; Magnúsdóttir et al., 2017; Pornputtapong, Nookaew, & Nielsen, 2015). This success has been partially derived from the improvement of automatic reconstruction algorithms and curating techniques (Baart & Martens, 2012; Dias, Rocha, Ferreira, & Rocha, 2015; Kim et al., 2012; Thiele & Palsson, 2010), which has enabled relatively fast reconstruction of metabolic models from existing metabolic databases like the KEGG database (Kanehisa et al., 2016) or BioCyc (Caspi et al., 2016). In addition, the agreement on a common language to display such models, the Systems Biology Markup Language (Hucka et al., 2003) has standardized the metabolic model reconstructions and facilitated their use and investigation by the scientific community.

The metabolic network in a GEM is represented by a directed weighted hypergraph, *H(M,R)*, in which the set of nodes correspond to the metabolites, *M*, and the set of hyperedges to the biochemical reactions, *R* (Figure 1.1A). Thereby *H* captures the topology of the metabolic network of a species. Further, *H* is encoded by the

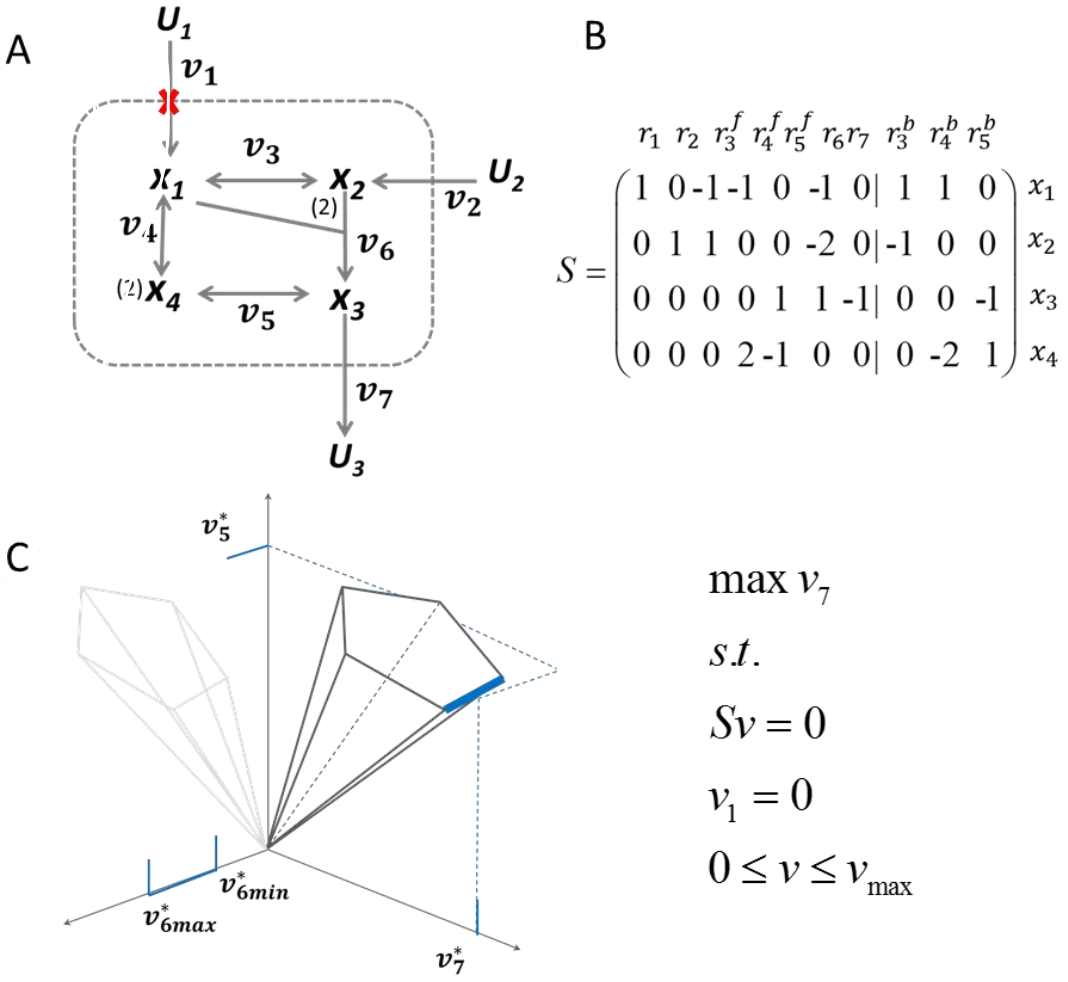


Figure 1.1. Cartoon example depicting the core idea behind constraint-based approaches and optimization methods. (A) A metabolic network can be represented as a hypergraph, where nodes, i.e., metabolites, are connected through hyperedges, i.e., chemical reactions transforming metabolites. Each metabolite participates in a reaction, either as a substrate or a product, with a specific stoichiometry. (B) The topology of the hypergraph, and the stoichiometry of the reactions, are captured by the stoichiometric matrix S. In S, column vectors depict the stoichiometric coefficients of each metabolite, x_i, in each reaction, r_i, where negative coefficients correspond to consumption and positive to production of a metabolite. Thermodynamically reversible reactions, r_{3-5}, may be split into two components, the forward, r^f, and backward, r^b, direction, which are explicitly included in an extended stoichiometric matrix. In this case, three extra columns are added—the backward direction for each reversible reaction—in which the signs of the stoichiometric coefficients are flipped. (C) The steady-state assumption and the flux bound constraints define a flux cone, which contains all feasible flux distributions. An optimization problem may be defined to find a particular flux distribution. In this example, we want to optimize the production of U_3, i.e., maximize v_7, provided that the system can only feed on U_2, i.e., $v_1 = 0$. Optimal flux values, depicted in blue, may be unique for some reactions,

but other reactions can adopt a range of optimal flux values without affecting the maximum value of the objective function. In this case, the optimal value of our objective reaction, v_7^*, is on a edge of the cone that lays along the dimension of the reaction v_6, thus v_6 can vary its flux at the optimum. This is the basic principle behind alternative optima encountered in constraint-based approaches (discussed in the main text).

so-called stoichiometric matrix, *S,* in which the $s_{m,r}$ entry contains the stoichiometric coefficient of the metabolite *m* participating in reaction *r*—by convention, a negative coefficient indicates metabolite consumption whereas a positive coefficient indicates production by the reaction. Therefore, each column in *S* represents a chemical reaction in the metabolism of a species (Figure 1.1B). In addition to the stoichiometry and the topology of the network, a metabolic model contains information about the maximum and minimum flux capacity, *i.e.*, upper and lower bounds of each reaction, which is derived from experimental evidence. In general, reaction flux values are non-negative quantities. However, it is customary to define negative lower bounds for thermodynamically reversible reactions. This is because the flux, *v*, through a reversible reaction is implicitly interpreted as the difference between the flux through the forward, v_{for}, and the backward, v_{back}, direction, *i.e.,* $v = v_{for} - v_{back}$, where both $v_{for}, v_{back} \geq 0$. Alternatively, it is sometimes more convenient to split reversible reactions and explicitly include the forward and backward reaction as individual columns in *S*. In this case, the signs of the stoichiometric coefficients corresponding to the backward reaction are flipped (Figure 1.1B).

1.3.2 Genome-scale metabolic model predictions

GEMs are primary tailored to generate large-scale metabolic predictions, which generally consist of assigning flux values to each reaction in the metabolic network. These flux values describe the *metabolic state* of the network, and are obtained as solutions to a linear system of equations in which the domain is restricted by some metabolic and physiological constraints (Palsson, 2006). These solutions are usually not unique, therefore in most settings an underlying optimization problem is employed to select particular solutions. During the optimization, a property of the metabolic network is usually optimized subject to the metabolic and physiological constraints (Orth, Thiele, & Palsson, 2010). For instance, a typical optimization problem consists of finding the metabolic state that maximizes the production of a particular metabolite by the network, subject to mass conservation and maximum flux capacity constraints, and when only some input molecules are allowed to enter the network (Figure 1.1C).

The combination of *constraint-based approaches* (CBAs) and optimization techniques render GEMs very useful frameworks to investigate metabolic function under different environmental scenarios. Importantly, GEMs also allow a mathematically precise mapping between genotype and phenotype. Here, genotype corresponds to the set of reactions in the network—which is identified to the set of

included enzymes. In contrast, phenotype corresponds to their flux value or, more coarsely, to their activity state: active if a reaction carries an absolute flux value greater than a small threshold, inactive otherwise. The genotype-to-phenotype mapping provided by GEMs allows a quantitative investigation of the phenotypic effects due to changes in the genotype, which currently constitutes a major open problem in biology (Dowell et al., 2010; Nuzhdin, Friesen, & McIntyre, 2012; Pigliucci, 2010). In the next sections we will review the mathematical formalism that makes these large-scale metabolic predictions possible.

1.3.3 The rationale of constraint-based approaches

A complete description of the state of a metabolic network—*i.e.,* the time evolution of metabolite concentrations and reaction fluxes—is encoded by the dynamical system

$$\dot{X}_i = \sum_{j \in I} s_{ij} \varphi_j(X(t)) - \sum_{k \in O} s_{ik} \varphi_k(X(t)), \; i = \{1, ..., m\} \tag{1.16}$$

where $\dot{X}_i$ represents the (first) time derivative of the concentration of metabolite *i,* which is computed as the difference between the weighted sum of the metabolic fluxes of the input— index set *I*—and output—index set *O*—reactions which produce and consume metabolite *i,* respectively. Each flux is weighted by the stoichiometric coefficient s_{ij}, encoded in the corresponding entry of *S*. The metabolic fluxes, or reaction rates, $v = \varphi(X)$, are, in general, non-linear functions of $X \in \mathbb{R}^m_{\geq 0}$, the concentrations of the metabolites in the network. Examples of common rate functions are the mass action or the Michaelis-Menten kinetics discussed in 1.2.

However, a genome-scale dynamical description of metabolism is currently unfeasible. This is mainly due to the lack of knowledge on the correct form of φ, particularly regarding the parametrization of the system. For instance, even when considering the simplest rate function, the mass action formalism of (1.4), we need to experimentally determine the values of the rate constants, *k,* for each reaction and under the studied physiological conditions. This situation is even more challenging when using the Michaelis-Menten formalism (1.14), since both the maximum velocity, v_{max}, and the affinity, $K_{M,}$ constants must be experimentally determined. Current experimental evidence only supports the parametrization of small-scale dynamical systems, which usually consist of at most, tens of metabolites and reactions. Therefore, the description of the metabolic state of large-scale networks requires different approaches.

CBAs came to rescue this situation soon after the first GEMs where reconstructed, and still constitute the only way to obtain genome-scale metabolic predictions. These approaches focus on reaction flux values and neglect the description of metabolite concentrations. Therefore, the system in (1.16) becomes

$$\dot{X}_i = \sum_{j \in I} s_{ij} v(t)_j - \sum_{j \in O} s_{ij} v(t)_j, \; i = \{1, ..., m\} \tag{1.17}$$

or in matrix notation

$$\dot{X} = Sv(t). \tag{1.18}$$

The system in (1.18) is a linear system in which the stoichiometric matrix S maps the vector of n reaction fluxes, $v \in \mathbb{R}^n$, to the vector of time derivatives of the m metabolite concentrations $\dot{X} \in \mathbb{R}^m$, where both v and $\dot{X}$ are unknown. If we further restrict our attention to the special case $\dot{X} = 0$—*i.e.*, when the system is at steady-state and thus the concentrations are constant—we then obtain

$$Sv = 0, \tag{1.19}$$

which is solvable for v through standard linear algebra techniques.

The system in (1.19) is usually underdetermined, since more reactions (columns in S) than metabolites (rows in S) are normally found in metabolic networks. Hence, solutions to (1.19) correspond to v that are compatible with a steady-state of the dynamical system in (1.16)—in this context, a particular reaction rate vector v is commonly known as a *flux distribution.* The steady-state assumption allows easy computation of flux distributions. However, this comes at the cost of losing information about the metabolite concentrations and the dynamics of the system in (1.16).

The steady-state assumption can be motivated from two perspectives. On the one hand, metabolic reactions occur at much shorter time-scales than that of other cellular processes, such as gene expression. This observation justifies treating metabolite concentrations as constant under the larger time-scales of measured physiological processes, an approach known as the *quasy-steady-state* approximation (Heinrich & Schuster, 1996; Varma & Palsson, 1994). On the other hand, on the long run metabolism must be mass balanced, *i.e.,* the production and consumption of most metabolites must be similar as to avoid unlimited accumulations or depletions (Reimers & Reimers, 2016)

Finally, the feasible space of (1.19) must be further restricted by physiological constraints, such as the previously discussed thermodynamic and flux capacity constraints. Thus, the actual feasible space is a solution to

$$\begin{aligned} & Sv = 0 \\ & s.t. \text{ (such that)} \\ & v_{\min} \le v \le v_{\max} \end{aligned} \tag{1.20}$$

The system of linear equality and inequality constraints in (1.20) defines a convex polyhedron, additionally, if $v_i \ge 0, \; \forall i$—*i.e.*, when reversible reactions are split—this

polyhedron corresponds to a convex polyhedral cone (Rockafellar, n.d.), usually designated as the *flux cone,*

$$K = \{v \in \mathbb{R}^n : Sv = 0, v_{\min} \leq v \leq v_{\max}\} \tag{1.21}$$

which contains all flux distributions, *v,* that are compatible with a steady-state of (1.16) and that satisfy the imposed thermodynamic and flux capacity constraints. The general goals of CBAs then reduce to: *i*) investigating global properties of *K,* with the objective of characterizing the entire set of allowed flux distributions of a GEM, and *ii*) find a particular $v \in K$ that optimizes some pre-stablished property—*e.g.,* maximize the production of a certain metabolite.

In the first case, the flux cone can be fully characterized by enumerating the set of *extreme rays* generating *K* (Figure 1.2). The set of extreme rays is a set of feasible flux distributions with the property that any other feasible flux distribution $v \in K$ can be expressed as a non-negative, or *conic,* combination of them. Translated to a more formal language and denoting the set of extreme rays $E = \{e_1, ..., e_k\}$, with $E \subset K$ — dismissing the trivial case where *S* contains a single reaction column and *E* = *K*— we have,

$$\forall v \in K,\ v = \sum_{i=1}^{k} \alpha_i e_i,\ \alpha_i \geq 0\,. \tag{1.22}$$

Additionally, extreme rays must be conically independent—*i.e.,* each $e \in E$ cannot be expressed as a conic combination of the rest of the members of *E.* Several variants of generating sets have been proposed besides extreme rays, such as *extreme pathways* (Wiback & Palsson, 2002) and *elementary flux modes* (Zanghellini, Ruckerbauer, Hanscho, & Jungreuthmayer, 2013). Although similar in concept, these variants differ in subtleties, mainly due to how reversible reactions are treated—an excellent review on this topic is provided in (Llaneras & Pic, 2010). In any case, the enumeration of extreme rays, extreme pathways or elementary flux modes is restricted to analyzing small- and medium-size subnetworks. The reason is that the number of extreme rays, elementary flux modes or extreme pathways scales up exponentially with the size of the network. This characteristic renders the enumeration computationally challenging or even impossible when using regular computers. Additionally, even if possible, the interpretation of such big number of generators—reaching hundreds of millions for a regular GEM—is cumbersome (Horvat, Koller, & Braunegg, 2015; Terzer, 2009).

An alternative way of investigating *K* relies on drawing random samples of flux distributions in *K,* and then analyzing general properties of the samples (Figure 1.2). This approach provides insights on the allowable flux values that each reaction in *v* can take such that $v \in K$. Specifically, the samples allow generating distributions of flux values for each reaction, which allows an indirect, but computationally affordable, characterization of the flux cone. Moreover, efficient algorithms permit

extracting large samples of flux distributions using regular computers, even for full-sized GEMs (De Martino, Mori, & Parisi, 2015; Schellenberger & Palsson, 2009). Thereby, random sampling constitutes a common approach to investigating general properties of *K*. Finally, we can also apply a *flux variability analysis* (R. Mahadevan & Schilling, 2003) to *K*. This approach computes the minimum and maximum flux bounds for each reaction, thereby complementing the random sampling approach by providing flux bounds to the sampled distributions (Figure 1.2). In chapters 3 and 4, we will see how both random sampling and flux variability analysis can be adapted to explore particular subsets of *K*, which are defined by including additional constraints in (1.20) and represent specific cellular scenarios derived from a general GEM.

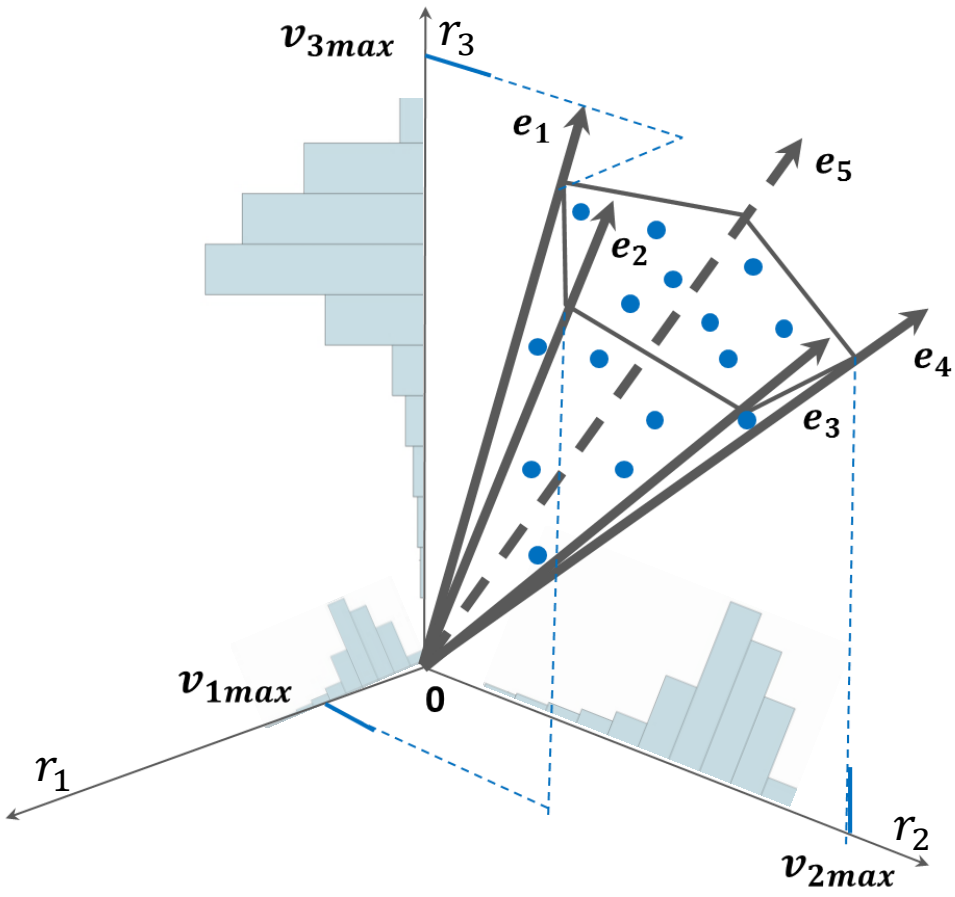


Figure 1.2. An illustration of three approaches employed to characterize the flux cone. *A full characterization of the flux cone, K, requires the enumeration of a generating set. In this cartoon example, the five extreme rays, e_1 to e_5, generating K correspond to the edges of the polyhedral cone. Any point, i.e., flux distribution, in K can be expressed as a conic, i.e., non-negative, combination of the extreme rays. A random sample of points, depicted as blue dots, is a more computationally tractable way of characterizing K. Once we have a sample of flux distributions, we can generate distributions of flux values for each of the reactions in the GEM, in this example only three reactions, r_1, r_2, r_3. Finally, flux variability analysis provides the minimum and maximum flux values for each reaction in the GEM. In the example, the three reactions range from a minimum flux value of zero to the corresponding maximum allowable flux value, v_{max}, which serve as a bound to the sampled distributions. The maximum flux value depends on the geometry of K, and on arbitrary maximum flux capacity constraints that truncate the cone.*

The second case—finding a particular $v \in K$ that optimizes a given metabolic property—is the most widely employed approach, and also was the starting point of CBAs, with the so-called *flux balance analysis* (FBA) (Orth et al., 2010). All methods in this category are based on numerically solving some sort of convex optimization problem. Optimization problems define an objective function which is optimized (*i.e.,* minimized or maximized) and which is defined over a restricted domain. In the case of CBAs, the restricted domain corresponds to *K*—or subsets of *K* if additional constraints are considered—while the objective function may take diverse forms. All computational methods proposed in this thesis rely on some kind of convex optimization problem. For this reason, we will briefly review some general characteristics of convex optimization problems, and then continue with a brief review of the main types of optimization problems employed in CBAs.

1.3.4 Brief notes on convex optimization

A general mathematical optimization problem consists of minimizing a real-valued function over a constrained or *feasible* domain (Boyd & Vandenberghe, 2010). This problem is formalized in general as,

$$\begin{aligned} &\min f(v) \\ &s.t.\ \ g_i(v) = b_i, \quad i = 1,...,p \\ &\qquad h_i(v) \le c_i, \quad i = 1,...,q \end{aligned} \tag{1.23}$$

where the objective function $f : \mathbb{R}^n \to \mathbb{R}$ is minimized subject to *p* equality and *q* inequality constraint functions $g_i, h_i : \mathbb{R}^n \to \mathbb{R}$, respectively, which define the feasible domain $F = \{v : g_i(v) = b_i, h_i(v) \le c_i\}$—note that we do not consider the maximization of *f* since it is equivalent to minimizing (–)*f*. A vector $v^* \in F$ is considered a *global* optimum if it minimizes the objective function over the entire feasible domain, that is, $\forall v \in F, f(v^*) \le f(v)$. In contrast, a vector w^* is considered a *local* optimum if it minimizes the objective function only over a δ-neighboring region $U = \{w : \|w - w^*\|_2 < \delta,\ \delta > 0\} \subset F$ centered on w^*, that is, $\forall w \in U, f(w^*) \le f(w)$. Optimization problems can be classified into groups based on the form of the objective and constraint functions (Boyd & Vandenberghe, 2010; Rockafellar, n.d.). A particular group is that of *convex* optimization problems, in which both, the objective function and the feasible domain defined by the constraints are convex. We designate a function *f* as convex if, $\forall v, w \in \operatorname{dom} f$, the following inequality holds

$$f(\alpha v + (1-\alpha)w) \le \alpha f(v) + (1-\alpha) f(w),\ \alpha \in [0,1], \tag{1.24}$$

additionally, if the strict inequality, <, holds in (1.24) then *f* is designated as *strictly* convex. On the other hand, a convex set, *F*, satisfies,

$$v, w \in F \Rightarrow \alpha v + (1-\alpha)w \in F,\ \alpha \in [0,1]. \tag{1.25}$$

In both cases, the concept of *convex combination* of two points, *v, w, i.e.,* $\alpha v + (1-\alpha)w$ with $\alpha \in [0,1]$, is crucial, and geometrically corresponds to the line segment connecting *v* and *w*. In fact, statement in (1.24) can be phrased as: a function is convex if its image in the line segment connecting any two points, *v, w*, in its domain lies below or on the line segment connecting the images of the function at these two points (Figure 1.3A). Statement (1.25) may be phrased: a set is convex if the line segment connecting any two points, *v, w,* in the set is also contained in the set (Figure 1.3B). These two properties grant convex optimization problems the following, and perhaps most important, characteristic: in convex optimization problems, any local optimum is also a global optimum, that is,

$$f(w^*) \leq f(w) \Rightarrow f(w^*) \leq f(v),\ \forall w \in U,\ \forall v \in F \tag{1.26}$$

Statement (1.26), a classic result in convex optimization (Boyd & Vandenberghe, 2010; Rockafellar, n.d.), can be easily proven in the following way. Suppose that we construct a point $z = \alpha v + (1-\alpha)w^*,\ \alpha \in [0,1]$ which is a convex combination of a local optimum w^* and a general point $v \in F$ no necessarily contained in *U*, then, since *F* is a convex set we have $z \in F$ by property (1.25). Now, as $\alpha \to 0$, we see that $z \to w^*$, hence $z \in U$ for a sufficiently small α. Since w^* is a local optimum, this implies that $f(w^*) \leq f(z)$. Substituting z and applying the convexity property in (1.24) we obtain

$$f(w^*) \leq f(\alpha v + (1-\alpha)w^*) \leq \alpha f(v) + (1-\alpha) f(w^*) \tag{1.27}$$

hence

$$f(w^*) \leq \alpha f(v) + (1-\alpha) f(w^*) \tag{1.28}$$

which implies

$$\alpha f(w^*) \leq \alpha f(v) \tag{1.29}$$

and

$$f(w^*) \leq f(v)\ \blacksquare \tag{1.30}$$

In practice, statement (1.26) implies that a suitable algorithm is guaranteed to find the global optimum of a convex optimization problem—provided that such optimum exists. This makes convex optimization problems especially amenable to solve, which

is far from being the case when dealing with non-convex optimization problems (Boyd & Vandenberghe, 2010).

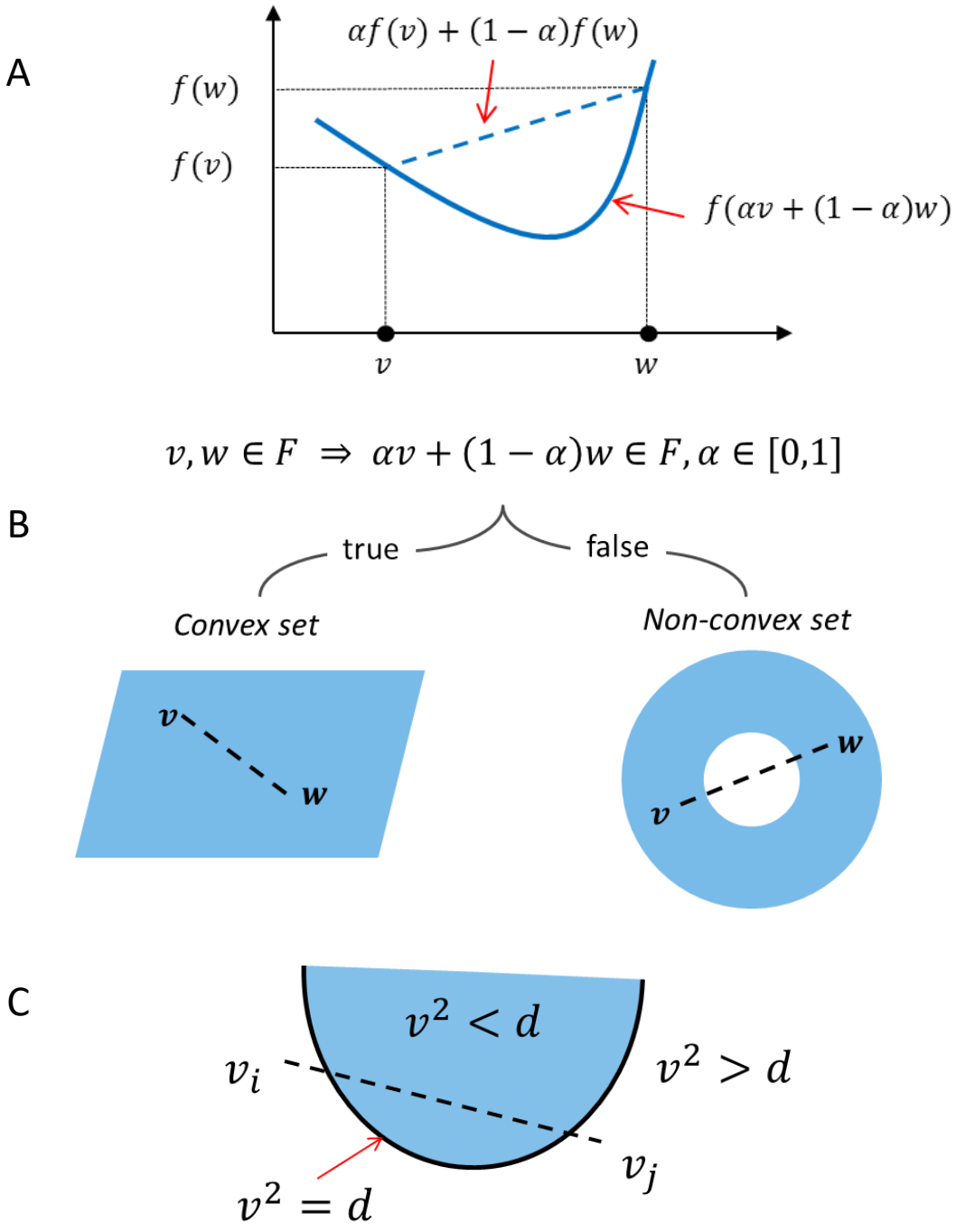


Figure 1.3. Illustrations of convex functions and convex sets. *(A) A function f is convex if the image of the function corresponding to the line segment between two points $v, w \in \mathrm{dom}\, f$ lies below or on the line segment connecting the images $f(v)$, $f(w)$ of the two points. Note that all affine functions, hence all linear functions, satisfy this condition since $f(\alpha v + (1-\alpha)w) = \alpha f(v) + (1-\alpha) f(w)$. (B) A set is convex if any line segment connecting two points in the set is also in the set, thus the 2D torus is non-convex in this example while the parallelogram is. (C) Quadratic equally constraints are never*

convex, since the points on the parabola $v^2 = d$ are not a convex set. In contrast, the set of points satisfying $v^2 < d$ form a convex set and hence quadratic inequality constraints of this form keep the convexity of the optimization problem. Note that the set of points satisfying $v^2 > d$ is again non-convex.

1.3.4.1 Types of convex optimization problems employed in constraint-based approaches

Convex optimization problems can be further classified in different categories based on the specific form of their objective and constraint functions. The simplest possible scenario is that of linear objective and constraint functions, in which case the optimization problem, usually designated as a *linear program* (LP)*,* takes the form,

$$\begin{aligned} &\min\ d^T v \\ &s.t.\ Av = b \\ &\quad Bv \le c \end{aligned} \tag{1.31}$$

where *d* is a vector of constant weights for the *n* variables in $v \in \mathbb{R}^n$ —the superscript ^{T} represents transposition—and the *p* equality and *q* inequality constraints are represented by matrices A^{pxn} and B^{qxn}, and the right-hand-side vectors *b* and *c*, respectively. Additional constraints can be imposed on the values that *v* can take, these constraints can take the form of lower (lb) or upper (ub) bounds on *v*, *i.e.*, $lb \le v \le ub$, or even constrain the totality or part of the variables in *v* to take integer values. In the last case, these optimization problems are known as *integer linear programs* (ILPs) and *mixed-integer linear programs* (MILPs), when all or part of the variables take only integer values, respectively.

FBA (Orth et al., 2010), the most widely employed CBA is formulated as a LP. In fact, we only need to let *A* = *S*, the stoichiometric matrix of the metabolic network, *b* = 0, the null vector, and dismiss the inequality constraints, excluding $lb \le v \le ub$, in (1.31), with the feasible space *F* = *K,* to arrive at the formulation of FBA. The objective function

$$f = d^T v = \sum_{i=1}^{n} d_i v_i \tag{1.32}$$

consists of a linear combination of reaction flux values which are weighted by real coefficients, d_i. In the first implementations, although still widely employed, f corresponded to a single reaction, the biomass reaction, which is an artificial metabolic reaction included in *S* to represent cellular growth. Specifically, the biomass reaction, v_{bio}, is a sink reaction—*i.e.,* its products are not consumed by any other reaction in *S*—which drains certain metabolites. These metabolites and their stoichiometry are experimentally defined, and are assumed building blocks to cellular growth (Feist & Palsson, 2010). In the LP of FBA, v_{bio} is then maximized—in (1.31)

this would correspond to minimizing $(-)d^T v$, where all entries in d are 0 minus the entry corresponding to the index of v_{bio} in S, which is 1. With this strategy, we thus seek for a $v \in F$ that maximizes v_{bio}, *i.e.*, we assume that a given species optimizes its metabolic state, the flux distribution at steady-state, to maximize cellular growth. FBA has expanded to include other definitions of f, such as the minimization of energy consumption (Savinell & Palsson, 1992) or the minimization of metabolic costs associated to operating reactions (Holzhütter, 2004).

Besides FBA, almost all remaining optimization problems that arise in CBAs are LPs or MILPs. However, in some applications *quadratic programs* (QPs) are also employed. In this case, the objective function is quadratic, which gives the general form,

$$\begin{aligned} &\min\ v^T Q v + d^T v \\ &s.t.\ Av = b \\ &\quad Bv \leq c \end{aligned} \tag{1.33}$$

where Q is a positive semi-definite (negative semi-definite in case of maximizing the objective) matrix of weights for the quadratic terms of the objective function. Additionally, a quadratic function may be combined with integer constraints to render *integer quadratic programs* (IQPs) or a *mixed integer quadratic programs* (MIQPs).

A natural implementation of a QP arises when the second norm of the difference between two vectors is minimized, *e.g.,*

$$f = \sum_{i=1}^{n} (v_i - w_i)^2 \tag{1.34}$$

for some $v, w \in \mathbb{R}^n$. For instance, the method of minimization of metabolic adjustment (MOMA) (Segrè, Vitkup, & Church, 2002), finds a feasible flux distribution $v \in K$ that minimizes the distance to a fixed $w \in K$ previously obtained through FBA. In this setting, w corresponds to an optimal flux distribution, *e.g.,* maximizing growth, of a wild-type specimen, while v corresponds to a mutant specimen, *e.g.,* a knock-out strain where some reactions in v have been constrained to carry zero flux. MOMA thus hypothesizes that mutant strains will reach a steady-state that is most similar to the wild type state. The last two cases, IQPs and MIQPs, seldom appear in CBAs, however, one of the methods developed and presented in this thesis, RegrEx (Robaina Estévez & Nikoloski, 2015), is formulated as a MIQP (Chapter 2).

All constraint functions presented so far have been linear. However, quadratic constraints may also be included in a convex optimization problem, in which case the optimization programs are known as *quadratically constrained linear programs* (QCLPs) or *quadratically constrained quadratic programs* (QCQPs). However, only

quadratic inequality constraints of the form $v^T Q_i v + r_i^T v \leq t_i$, where Q_i must be a positive semi-definite matrix, are valid to guarantee that the feasible space is convex (Figure 1.3C). Quadratic equality constraints may arise naturally in certain optimization problems. In Chapter 3, we will encounter quadratic equality constraints arising during the evaluation of the alternative optima space of RegrEx. In this case, a change in the formulation of RegrEx is required to eliminate the quadratic equality constraints, hence exploiting convex optimization to evaluate its alternative optima space.

1.4 Context-specific metabolic predictions

The metabolic networks contained in GEMs capture all known biochemical reactions of a species. This implies that all biochemical reactions are considered when using GEMs to obtain metabolic predictions. However, metabolic networks can specialize in certain metabolic functions, which only require a subset of the reactions included in the GEM. For instance, this may be the case of a bacterium like *Escherichia coli* growing in aerobic or anaerobic conditions. Certainly, after reaching a steady-state, the metabolic network will focus on operating aerobic metabolic pathways in the first case, such as the respiratory chain, or anaerobic pathways, such as fermentation pathways, in the second. A richer metabolic specialization is found among multicellular organisms. There, a multitude of cell-types specialize in different metabolic functions while deliberately obliterate others (Figure 1.4). Therefore, to render more realistic predictions, the *context-specific* specialization of metabolism must be reflected in GEMs when simulating scenarios in which the metabolic context is important—*e.g.*, when dealing with different cellular types in multicellular species.

The advent of the first GEMs of multicellular species stimulated the development of computational methods to obtain context-specific metabolic predictions. The general idea in these methods is simple: use context-specific experimental data to reduce the solution space of (1.20), leaving flux distributions $v \in K$ that are also compatible with data. This context-specific solution space can then be directly employed to predict flux distributions, *e.g.,* (Colijn et al., 2009; D. Lee et al., 2012; Recht et al., 2014). However, more commonly, the context-specific flux distributions are employed to determine which reactions are likely to be active, *i.e.,* carry non-zero flux, under the particular context, *e.g.,* (Jerby, Shlomi, & Ruppin, 2010; Schultz & Qutub, 2016; Shlomi, Cabili, Herrgård, Palsson, & Ruppin, 2008; Vlassis, Pacheco, & Sauter, 2014; Yuliang Wang, Eddy, & Price, 2012). This information is then used to prune the genome-scale or *generic* network in the GEM, as to extract a context-specific (sub)network, which preferentially includes the set of active reactions. The different data fields in the GEM—such as the stoichiometric matrix, the flux capacity bounds and the lists of reaction, metabolite and gene names—are then updated to obtain a context-specific metabolic model, which can be interrogated with the usual CBAs (Figure 1.4). In the following, we will refer to both, context-specific flux distributions

and context-specific models, with the general term *context-specific metabolic predictions.*

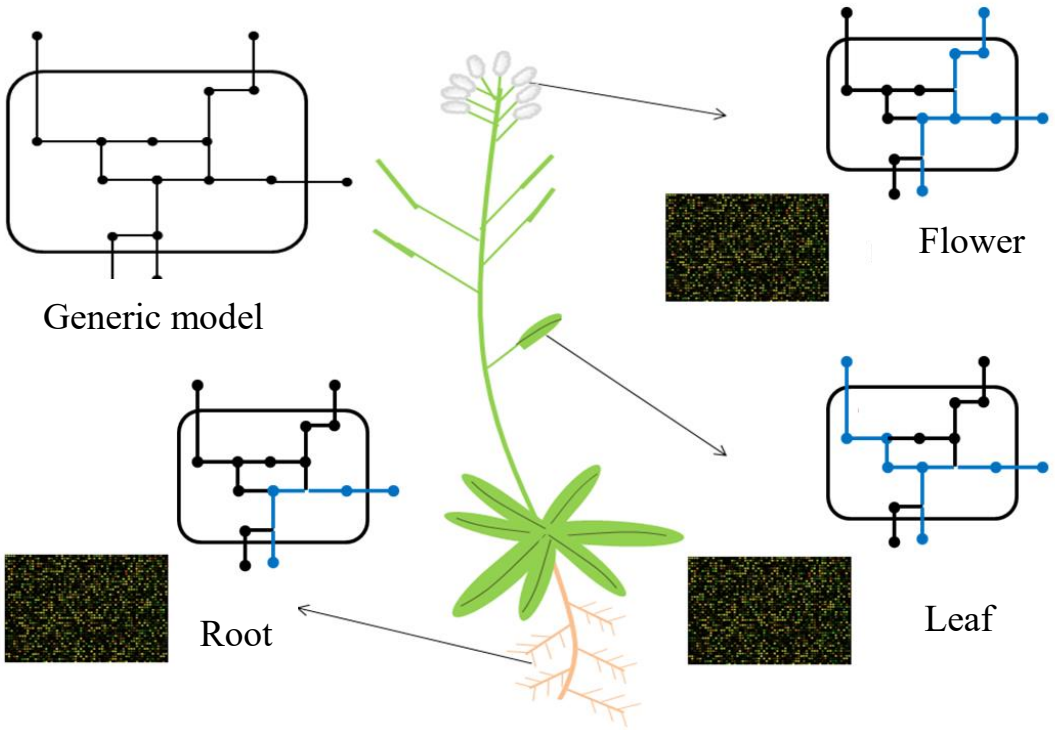


Figure 1.4. Metabolic activity is context-specific. *A genome-scale model contains the entirety of known metabolic reactions of a species, such as the plant* Arabidopsis thaliana*. All cells in* A. thaliana *contain the same genome. Therefore, they can potentially realize all reactions contained in the metabolism of the species. However, tissues specialize in certain metabolic tasks, rendering metabolism tissue-specific. A more realistic model of the metabolic processes must take into account this specialization. To address this, context-specific models can be reconstructed from a genome-scale or generic model and experimental data. The experimental data are employed to prune the generic model, selecting reactions that are most likely active under the context. In this sense, experimental data define the context while the genome-scale model guarantees that the selected reactions will still be mass balanced, i.e. supporting a steady-state, and will satisfy the flux capacity constraints. In this example, three organ-specific models have been extracted after integrating gene expression data of flower, leaf and root. Higher resolutions, e.g., cell-specific models, can be obtained if experimental data are available.*

1.4.1 Main types of experimental data integrated in genome-scale models

Diverse kinds of experimental evidence can be integrated into GEMs to obtain context-specific predictions. Metabolomics data may be used to determine which metabolites are being synthetized in a given context, and constraint model predictions to guarantee their production (Schmidt et al., 2013). Alternatively, metabolomics data can be employed to constraint the maximum flux values of some reactions in a GEM (Töpfer, Kleessen, & Nikoloski, 2015). Proteomics and transcriptomics data serve as a proxy for reaction activity—based on the assumption that reaction flux is controlled by enzyme levels—although proteomics data provide a more direct approximation, since directly quantify enzyme levels. Some studies have employed existing proteomics data to reconstruct tissue-specific human models (Agren et al., 2012, 2014), benefiting from one of the largest proteomics databases, the Human Protein Atlas (Marx, 2014). However, the majority of methods rely on transcriptomics data to obtain context-specific predictions for two main reasons. Firstly, current techniques allow cheap and fast transcript quantification, while measuring protein levels require more costly and harder to implement techniques. Thus, in most cases proteomics data are not available. Secondly, the coverage—*i.e.,* the number of reactions for which experimental data can be associated—of transcriptomics data is larger than that of proteomics. Thus, in most cases, proteomics data have to be supplemented with transcriptomics data to cover the missing reactions (Agren et al., 2012).

As commented before, measurements of transcript or protein levels can only serve as a proxy of the activity of metabolic reactions, which ultimately depends on other factors, such as metabolite concentrations and regulatory processes. However, their usage is justified by the fact that, to date, they are the only data type with a genome-scale coverage. This issue will be further discussed in section 4.1. On the other hand, measurements of metabolic fluxes provide the most direct evidence of reaction activity. Yet, although recent advances in experimental techniques are promising (Vinaixa et al., 2017), a genome-scale quantification of metabolic fluxes is still unfeasible, thus making this data type insufficient in most settings. We will next review two widely employed experimental techniques to generate transcriptomics data, their advantages and disadvantages, and the main methods to map these data into the reactions of a GEM.

1.4.1.1 DNA microarrays

DNA microarrays arose in the early 1990s, and were the first available tool to quantify cellular mRNA levels at a genomic scale in a single experiment (Butte, 2002). A microarray consists of a collection of ordered microspots on a solid surface (like glass or plastic). Each microspot contains thousands of DNA fragments, *i.e.,* probes, representing different genomic features, such as gene fragments. To measure transcript levels in a microarray experiment, total RNA is extracted from the

biological sample, converted to complementary DNA (cDNA), labelled with a fluorophore, and exposed to the microarray as to allow hybridization between the probes and their cDNAs. Microarrays are finally passed through a washing process to eliminate spurious hybridizations, followed by an optical reading of the fluorescence intensity of the fluorophores under the right emission wavelength—microarrays are exposed to the appropriate absorption wavelength of the particular fluorophore used. The logic behind this quantitative method is that fluorescence intensity is proportional to the number of cDNA molecules hybridized on each microspot. Raw intensity data must be processed to allow a biologically meaningful interpretation of the results, which is especially important when comparing gene expression among different samples. Data processing includes correcting for background—*i.e.,* unspecific—hybridization, as well as normalizing the intensities between different samples, as to account for sample-specific differences in the hybridization conditions. Popular methods for background correction include subtracting the average inter-microspot signal to all microspot entries, or subtracting a probe background signal specific to each microspot (Butte, 2002). In the last case, DNA microarrays include two classes of probes: match probes, with the original sequence, and mismatch probes, which contain a degenerate sequence. The signal corresponding to mismatch probes is taken as proportional to the intensity of unspecific hybridization of each probe in the microarray. Algorithms like MAS5 (Hubbell, Liu, & Mei, 2002) take into account both match and mismatch probes to subtract the background noise. In the case of multiple comparisons across different experiments, algorithms like the Robust Multi-Array Average (RMA) (Irizarry et al., 2003) allow to correct sample-specific differences in hybridization conditions as well as total RNA levels, thus rendering comparable results.

1.4.1.2 RNA-Seq

The RNA-Seq technique—from RNA sequencing—first appeared in 1998 (Z. Wang et al., 2009), and has been gaining popularity ever since. In contrast to DNA microarrays, RNA-Seq does not rely on DNA hybridization to identify specific transcripts, instead, it relies on their nucleotide sequence. In a RNA-Seq experiment, total RNA is also first extracted, however, at this step mRNA is enriched, either by using a specific tag (*e.g.,* poly-T), or by depleting ribosomal RNA levels. This step is required to increase the signal to noise ratio, since most of the cellular RNA corresponds to ribosomal RNA. After the enrichment step, cDNA is synthetized and the fragments are sequenced by next generation sequencing techniques (Z. Wang et al., 2009). The sequenced fragments, called reads, are then usually mapped to a given genome, although some techniques allow for a *de novo* assembly of reads (Haas et al., 2013). The assumption here is that gene expression levels are proportional to the number of mapped reads to the specific locus to which a gene is associated. As in the case of DNA microarrays, raw RNA-Seq data must be processed to render biologically meaningful results. In particular, read counts must be normalized by the coverage or depth of the sequencing step, which may differ between experiments.

Therefore, read or fragment[2] counts are usually measured in reads or fragments per million mapped reads, RPM and FPM, respectively. Additionally, since gene lengths vary and longer genes tend to produce more reads, the final results are usually given in fragments or reads per kilobase of transcript per million mapped reads, FPKM and RPKM, respectively.

1.4.1.3 Drawbacks and advantages of DNA microarrays and RNA-Seq

RNA-Seq is currently the preferred technique to measure gene expression levels. There are a number of advantages to using RNA-Seq instead of DNA microarrays. In first place, results are more reliable since sequencing is more precise than hybridization to identify DNA fragments. Secondly, RNA-Seq has a wider dynamic range (Zhao, Fung-Leung, Bittner, Ngo, & Liu, 2014), free of probe saturation issues and more sensitive to small expression levels, since, in theory, a single fragment in the sample is enough for a gene to be detected. In the case of DNA microarrays, small concentrations of a specific transcript may not be detected due to the background noise. Thirdly, RNA-Seq renders gene expression values that are absolute. In contrast, microarray data render relative values due to the probe effect (C. Chen et al., 2011). Finally, RNA-Seq allows more flexible measurements of the transcriptome, since it can take into account non-coding RNAs, alternative splicing and virtually any RNA fragments that can be mapped onto a given genome. Altogether, these advantages of RNA-Seq over DNA microarray render transcript levels that are more in accord to measure protein levels from a same experiment (Zhao et al., 2014), and thus justify its used as a default technique to quantify gene expression. However, besides these advantages, DNA microarrays are still widely used. The main reasons are threefold: *i)* DNA microarrays are easier and faster to use, due to efficient commercial implementations, *ii*) they are cheaper and thus more accessible to researchers and *iii*) large resources of microarray data are freely accesible online, such as the GEO database (Barrett et al., 2013), which is not yet the case for RNA-Seq data.

1.4.2 Mapping data into genome-scale models

In CBAs, the metabolic state of a GEM is described by a flux distribution at steady-state. However, we have seen that the experimental data employed to obtain context-specific metabolic predictions do not directly quantify metabolic fluxes. Instead, indirect measures of metabolic activity, such as transcript and to a lesser extend protein levels are used. These data are sometimes complemented with metabolite levels, when available, or even with a variety of biochemical information—including enzyme activity essays and immunostaining—that is retrieved from biochemical databases (Caspi et al., 2016; Haug et al., 2013; Moretti et al., 2016; Placzek et al.,

[2] A fragment is a mapped region in the DNA that can contain more than one read.

2017). Therefore, experimental data have to be first associated, or mapped, to the reactions in the GEM, before employing it to obtain context-specific predictions.

Data types associated to gene or protein expression can be mapped to the reactions in a GEM through the gene-protein-reaction (GPR) rules, which are contained in the model. The GPR rules are logical rules that encode the causal relationships between genes and catalyzed reactions. The rules capture all the known enzymes catalyzing each reaction—*e.g.*, enzyme isoforms—as well as the genes coding for the enzymes. Genes that code for the same enzyme are related by an *AND* operator, while groups of genes that code for different enzymes catalyzing the same reaction are related by an *OR* operator. For example, the following rule:

$$\left(g_1 AND\ g_2\right)\ OR\ \left(g_1 AND\ g_3\right)\ OR\ \left(g_4 AND\ g_5\right) \tag{1.35}$$

associates five different genes to a hypothetical reaction. In this case, three different dimeric enzymes are able to catalyze the reaction; one enzyme is coded by genes g_1 and g_2, the other one by g_1 and g_3 and the last one by g_4 and g_5—where g_i represents a string containing the canonical gene identifier for a given species. Once GPR rules are evaluated, an expression value is associated to the reaction from the individual transcript levels of the genes in the rule. The most common way of doing this is by evaluating the *AND* relation as the minimum value between the two related genes, and the *OR* relation as the maximum value. For instance, if we assign a transcript level t_i to each gene g_i in (1.35), then the expression value, r, associated to the corresponding reaction would be computed as

$$r = \max(\min(t_1, t_2), \min(t_1, t_3), \min(t_4, t_5)) . \tag{1.36}$$

Alternatively, the *OR* relation can be evaluated as the sum instead of the maximum value between the two transcript levels (D. Lee et al., 2012), although this form is less commonly used.

The reasoning behind this mapping is the following: if a set of genes is required to code a given enzyme, then the gene with the smaller transcript level will limit the total expression of the enzyme—leaving aside other limiting factors downstream of gene expression, such as the stability of the mRNA. On the other hand, if two enzyme isoforms are able to catalyze the reaction, then the enzyme whose coding genes are preferentially expressed will be the largest contributor to the reaction flux among the two isoforms. If the sum instead of the maximum value is selected in the *OR* relation, then the evaluation is interpreted as the total catalytic contribution of the two isoforms. Additionally, protein levels can be mapped to individual reactions following the same rules, although genes must first be mapped to enzymes.

Other data types, besides transcript and protein levels, require a different mapping to the corresponding reactions. We already described two common ways of integrating metabolite data: forcing the inclusion of reactions producing a metabolite known to be present in a context (Schmidt et al., 2013), and using metabolite levels to directly

constraint the maximum flux capacity of some reactions (Töpfer et al., 2015). On the other hand, heuristic approaches may be used when dealing with heterogeneous data types, such as using enzyme activity or immunostaining essays together with transcriptomic, proteomic or metabolomic data. In this case, a total score is usually assigned to each reaction based on the contributions of each kind of experimental evidence (Wang et al., 2012).

The definition of such a heuristic approach may be difficult. Especially since it is not clear *a priori* how to weight the contributions of the different data types to the total score.. On the other hand, this approach allows integrating diverse sorts of experimental observations: from transcript and protein levels, to metabolomics, enzyme activity essays or, in general, any kind of available biochemical information—the term *bibliomics* data is usually employed in this case.

1.4.3 A classification of existing methods

In this section, we briefly review the main characteristics of the existing methods particularly developed to extract context-specific metabolic models. A more detailed mathematical classification of these methods is provided in Robaina Estévez & Nikoloski (2014).

Existing methods devoted to extract context-specific metabolic models differ in several aspects. The way in which data are used during the optimization, the number of additional metabolic functions that are imposed and the level of automation of the algorithms are some of them. Yet, they can be grouped into three main families of methods that follow similar optimization principles: the *GIMME-like*, *iMAT-like* and *MBA-like* families, named after the first method of each family in chronological order (Figure 1.5).

The GIMME-like and iMAT-like families are characterized by using optimization problems that maximize the similarity between a feasible flux distribution and the employed data. However, the main difference between methods in the two families is whether an additional required metabolic functionality is imposed during the optimization. Methods such as GIMME (Becker & Palsson, 2008) and GIM^{3}E (Schmidt et al., 2013) require, for instance, a minimum growth rate that must be satisfied by the optimal—*i.e.*, most similar to data—flux distribution; although any of the metabolic functions defined in FBA may be used. This additional constraint guarantees that the final context-specific models maintain a given metabolic operability.

On the other hand, methods such as iMAT (Shlomi et al., 2008) and INIT (Agren et al., 2012), in the iMAT-like family, do not impose any required metabolic functionality, and thus focus on maximizing the similarity to data—an updated version of INIT, tINIT (Agren et al., 2014) allows defining certain metabolic tasks that must be fulfilled by the context-specific model. However, the user still must

select an arbitrary threshold to binarize the employed data; hence these approaches are not totally unbiased.

The MBA-family, with MBA (Jerby et al., 2010), FastCORE (Vlassis, Pacheco, et al., 2014), FastCORMICS (Maria Pires Pacheco et al., 2015), mCADRE (Wang et al., 2012) and CORDA (Schultz & Qutub, 2016) as representatives, substantially differs from the previous. In this case, experimental data are used to define a core set of reactions that must be included in the final context-specific model. Importantly, the core set is stablished through a heuristic approach, for instance, by defining a threshold value to binarize the data associated to the reactions. For example, FastCORE uses the barcode algorithm (McCall et al., 2014) to binarize the original data. In more elaborated cases, the methods establish a score weighting the importance of each reaction based on different kinds of experimental data. In this latter case, mCADRE first classifies reactions into core and non-core based on the barcode algorithm. However, core reactions are then ranked in ascending order of specificity based on three criteria, which involve both, properties of the expression data and of the underlying metabolic network (Yuliang Wang et al., 2012). On the other hand, CORDA employs the discrete level of expression of the associated enzymes provided by the Human Protein Atlas (Uhlén et al., 2005; Uhlén, 2015) to classify reactions into four categories. In turn, the Human Protein Atlas assigns the discrete expression values (low, medium and high) based on an established procedure which weights the relative importance of the experimental evidence (Uhlén et al., 2005)

The approaches followed by the three families of methods offer different advantages. On the one hand, GIMME-like methods endow extracted models with a given context-specific metabolic operability—*i.e.,* a metabolic function that the model is able to perform—selected by the user prior to the extraction. Importantly, the metabolic function in this case is not merely achieved by the inclusion of a set of reactions performing it, as it could be done in MBA-like methods by including them in the core set. On the contrary, GIMME-like methods impose a minimum flux value, *i.e.,* greater than the arbitrary small threshold employed by MBA-like methods, to this set of reactions. Hence, GIMME-like methods may be a good choice when a metabolic function is known to operate under a particular context.

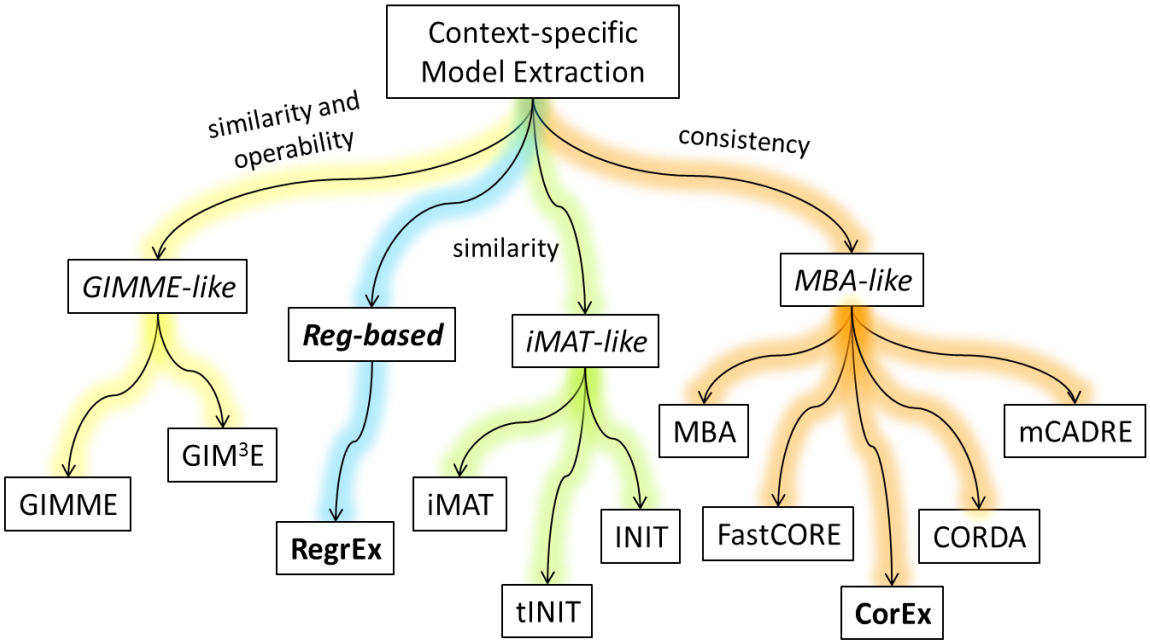


Figure 1.5. A classification of methods to extract context-specific metabolic models. *Different methods are able to extract context-specific models from genome-scale models and experimental data. These methods can be grouped into three main families based on their general formulation. GIMME-like methods aim at extracting models that are most similar to data and, at the same time, are able to conduct a pre-defined metabolic function, thus the reconstruction process is not fully data-driven. iMAT-like methods also aim at extracting models that are most similar to data, here, however, no assumptions about metabolic functionality are made, thus rendering the extraction fully data-driven. While both, GIMME-like and iMAT-like provide both a flux distribution and a context-specific model, MBA-like methods focus on reconstructing context-specific models. Moreover, they are capable of integrating diverse kinds of experimental data, since data are integrate in a semi-automatic, heuristic approach which precedes the optimization process. In this thesis, we will introduce a new family of methods, the Reg-based, for regularization, family, which has a unique member: RegrEx. RegrEx aims at obtaining models in a fully data-driven way, like the iMAT-like family. However, in contrast to the iMAT-like family, RegrEx does not require user-defined parameters, such as threshold values for the data. RegrEx finds the best parameter value during the optimization process, based solely on data, which renders the extraction process fully data-driven and automatic.*

On the other hand, iMAT-like methods may be better suited in contexts for which no clear knowledge about the metabolic function exists, or if some functions are known to operate but their relative importance is unknown. In such cases, it is best to follow a data-driven approach in which experimental data alone shape the context-specific network. Finally, MBA-like methods stand out for their flexibility when it comes to integrating different experimental data types. This is due to the heuristic approach followed during the definition of the core set, which is done prior to solving the optimization problem. This characteristic allows employing bibliomics data (section 1.4.2), which may be advantageous in situations where diverse experimental data types are accessible.

All the previous methods require some degree of biased intervention by the user. For instance, GIMME-like methods require not only the definition of a metabolic functionality but also the selection of a cost function to measure the discordance between model predictions and the employed data. Similarly, iMAT-like methods require selecting a threshold or penalty value to categorize the experimental data as context-specific—*e.g.,* establishing a threshold value for transcript levels as to filter out context-specific gene expression from background levels. Furthermore, MBA-methods require the establishment of the heuristic approach to translate all the employed, perhaps qualitatively diverse, experimental observations into a core set of reactions that must be active in the particular context. In all three cases, the election of arbitrary thresholds or heuristic approaches is not a trivial task—especially when little is known about a metabolic context. Additionally, slight variations in the threshold values or in the heuristic approaches can propagate and widely affect the end result of the context-specific network extraction.

An alternative method, operating in complete autonomy from user-based decisions and relying on a fully data-driven extraction, could be of help to this problem. RegrEx, which is presented in Chapter 2, fits this last description, and aims at extracting context-specific metabolic networks in a fully automated, data-driven process. As we will see in detail in Chapter 2, RegrEx (standing for Regularized Extraction) maximizes the similarity to data, thus similar to the iMAT-like family, to obtain both context-specific metabolic models and fluxes. However, the main characteristic of RegrEx is the usage of regularization[3] which allows a data-driven selection of the set of reactions that best capture the context-specific metabolic context. Moreover, the only free parameter used by RegrEx is optimized to maximize the overall similarity to data, as measured by the Pearson correlation, which is again a data-driven process.

1.4.4 Alternative optimal solutions in convex optimization

In section 1.3.4, we proved that every local optimum is also a global optimum in convex optimization problems. Now, in addition to this, we can say that if f is strictly convex near the optimum, then a unique vector v^* renders the global optimum $f(v^*)$ (Boyd & Vandenberghe, 2010). In fact, if we derive the previous proof (1.27-1.30) of statement (1.26) using the strict inequality we obtain $f(w^*) < f(v)$, which implies that y^* is the only global optimum—any other vector v renders a higher objective function value. If f is not strictly convex near the optimum, then we obtain again $f(w^*) \leq f(v)$, which opens the possibility to $f(w^*) = f(v)$ with $w^* \neq v$, that is, we may find a multitude of *alternative optimal* solutions to the convex optimization problem. These

[3] Regularization techniques are widely employed when estimating models from high-dimensional and redundant data sets, as a means to obtain sparser and simpler models. (Vidaurre et al., 2013)

alternative optimal solutions are all, by definition, global optima to the problem. However, they may substantially differ, *i.e.,* they can be located in distant regions of the feasible space, F. When solving a convex optimization problem, current algorithms commonly report only one optimal solution, even in cases where an alternative optimal set of solutions exists. This situation is not of major concern when the prime interest behind an optimization problem is finding an optimal solution, or the optimal value of the objective function, or both. However, if the particular values of the optimum v^* are of interest, then a complete answer to the convex optimization problem requires an evaluation of the alternative optimal set of solutions, as well as a decision procedure to select a representative solution vector from this set.

Alternative optima normally arise when solving FBA problems. In fact, the previously introduced *flux variability analysis* (R. Mahadevan & Schilling, 2003) was developed to analyze the set of alternative optimal flux distributions of FBA. The idea was to compute the minimum and maximum flux value of each reaction in the GEM such that the flux through the objective reaction is fixed to the (previously found) optimal value. The added constraint on the objective reaction effectively changes the shape of the flux cone. Therefore, investigating alternative optima is equivalent to exploring a modified flux cone, in which all flux distributions are optimal to a given optimization problem. Although alternative optima are well recognized in FBA problems (Kelk, Olivier, Stougie, & Bruggeman, 2012; S. Lee, Phalakornkule, Domach, & Grossmann, 2000; R. Mahadevan & Schilling, 2003; Müller & Bockmayr, 2014; Reed & Palsson, 2004), other settings lack of this recognition, and demand efficient methods to investigate alternative optimal solutions. For instance, this is the case when using CBAs to reconstruct context-specific metabolic models

1.4.5 Alternative optima in context-specific metabolic predictions

Existing methods to obtain context-specific metabolic predictions rely on solving some kind of convex optimization problem. For instance, the GIMME-like and the MBA-like families require solving LPs, while the iMAT-like family relies on MILPs—RegrEx solves either a MIQP or a MILP depending on the formulation, details will be given in Chapter 2. We have seen that, in general, the convexity of these optimization problems guarantees that a global optimum will be found by a suitable solver, although a multitude of alternative solutions may all be global optima. In our particular case, the existence of alternative optima translates to encountering a multitude of context-specific metabolic networks, or flux distributions, that are all equally compatible with the employed experimental data. Further, these alternative context-specific networks may differ substantially, *i.e.*, some context-specific reactions may not be consistently included among the alternative networks. Therefore, alternative optimal networks introduce ambiguity in context-specific metabolic predictions, and demand a careful evaluation of the alternative optima space to avoid misleading results. However, few studies have proposed any means to evaluate the alternative optima of a context-specific network extraction (Recht et al., 2014;

Rossell, Huynen, & Notebaart, 2013; Shlomi et al., 2008), and generally, they do not provide a quantitative analysis of the impact that the alternative optimal networks may have in further context-specific metabolic predictions. Thereby, this is currently an open field of research that must be addressed to guarantee that future context-specific metabolic predictions are robust.

1.5 Thesis outline

This thesis is the result of three scientific studies in which I have been the first co-author, the next chapters will present these studies. The first two chapters present a set of computational methods developed to obtain context-specific metabolic predictions. In contrast, the third chapter consists of an application of these methods to study the central carbon metabolism of a given cell type, the guard cells of *Arabidopsis thaliana.* Each chapter follows the structure of a scientific publication: a general introduction to the topic as well as the motivation of the study, followed by a "results & discussion" section, a general conclusion and outlook, and ending with the methods section. In the following, we briefly summarize the motivation and main findings of each chapter.

GEMs provide a means to investigate metabolism at a systems level. Recent computational developments provide context-specific metabolic predictions, which additionally allow investigating the context-dependent specialization of metabolism. Generally, these methods require a GEM and context-specific experimental data, such as transcript or protein levels, to extract a context-specific model. Additionally, the methods require the specification of either free parameters or a heuristic method to calibrate the influence of experimental data during the model extraction. For instance, the cutoff expression value required by GIMME (Becker & Palsson, 2008), GIM^{3}E (Schmidt et al., 2013) and iMAT (Shlomi et al., 2008), or the weights assigned to reaction-associated enzymes used by INIT (Agren et al., 2012) and CORDA (Schultz & Qutub, 2016), and utterly derived from a heuristic approach followed by the Human Protein Atlas database (Uhlén et al., 2005). On the other hand, as discussed in section 1.4.3, FastCORE (Vlassis, Pacheco, et al., 2014), FastCORMICS (Maria Pires Pacheco et al., 2015) and mCADRE (Yuliang Wang et al., 2012) rely on the cutoff value assigned by the barcode algorithm to pre-classify reactions into the groups of core (which should be included in the final model) and non-core reactions—mCADRE goes further and ranks reactions in the core following three different criteria.

These procedures are justified in each particular setting. However,t the nature of the final context-specific models depends on the election of particular threshold values or heuristics to pre-classifiy reactions (or the associated data value) into specific or non-specific to the context of interest. To address this issue, we developed the RegrEx (Regularized Extraction) method.Specifically, RegrEx exploits a regularization technique that requires a single parameter to extract a context-specific model. In turn,

the value of this parameter is automatically determined through a data-driven process. Morever, no cutoff value or heuristic approach is employed to pre-evaluate the specificity of reactions. Instead, reactions are selected through an optimization problem that uses the original (continuous) data values, which are subject to the constraints imposed by the underlying metabolic network.

Chapter 2 is dedicated to the RegrEx method, including a detailed description of its mathematical formulation and performance evaluation. In the performance evaluation RegrEx is employed to extract 11 organ-specific human models and compared with other contending methods. Additionally, the metabolic capabilities of the generated models are tested. Moreover, RegrEx predictions are validated with an independent data set, using human proteomics data instead of the transcript profiles employed to generate the models.

Chapter 3 addresses the issue of alternative optima in context-specific metabolic predictions. To this end, several computational approaches intended to analyze the alternative optima of context-specific extraction methods are provided. Our motivation started with analyzing the alternative optima space of RegrEx, for which we developed two algorithms, and proposed the Shannon entropy as a measure of the uncertainty generated by the alternative optimal solutions. However, we also developed a computational framework to analyze the alternative optima space of a different class of methods—the MBA-like family discussed in 1.4.3. This framework allows generating a sample of alternative optimal context-specific networks which can then be analyzed. These tools can be employed to reconstruct a consensus context-specific model, in which reactions with a higher representation in the alternative optima space are preferentially included. Therefore, our contribution paves the way towards obtaining more robust context-specific metabolic predictions.

Chapter 4 provides a specific application of the computational tools developed in Chapters 2 and 3. Specifically, we applied RegrEx to obtain context-specific metabolic predictions of the central metabolism of guard and mesophyll cells of *Arabidopsis thaliana.* By exploring the alternative optima space in each case, we could make robust comparisons between the two cellular scenarios. This comparison allowed the unravelling of specialized metabolic pathways in the guard cells, a system that is currently not very well understood. Additionally, an independent ^{13}C labelling experiment supported the key predictions of our modeling framework. Therefore, Chapter 4 serves as a "real life" example of the utility of the previously presented computational methods.

The last chapter of this thesis (Chapter 5) consist of a general discussion and outlook of the thesis. We will revisit Chapters 2 to Chapter 4 and discuss drawbacks and open problems in each case, as well as possible future directions to address them.

Chapter 2

Context-specific metabolic predictions with RegrEx

Published as:

Context-specific metabolic model extraction based on regularized least squares optimization

Semidán Robaina Estévez, Zoran Nikoloski

Systems Biology and Mathematical Modeling Group, Max Planck Institute of Molecular Plant Physiology, Potsdam-Golm, Germany

PLOS ONE (2015), DOI: 10.1371/journal.pone.0131875

Abstract

Genome-scale metabolic models have proven highly valuable in investigating cell physiology. Recent advances include the development of methods to extract context-specific models capable of describing metabolism under more specific scenarios (*e.g.,* cell types). Yet, none of the existing computational approaches allows for a fully automated model extraction and determination of a flux distribution independent of user-defined parameters. Here we present RegrEx, a fully automated approach that relies solely on context-specific data and ℓ_1-norm regularization to extract a context-specific model and to provide a flux distribution that maximizes its correlation to data. Moreover, the publically available implementation of RegrEx was used to extract 11 context-specific human models using publicly available RNA-Seq expression profiles, Recon1 and also Recon2, the most recent human metabolic model. The comparison of the performance of RegrEx and its contending alternatives demonstrates that the proposed method extracts models for which both the structure, *i.e.*, reactions included, and the flux distributions are in concordance with the employed data. These findings are supported by validation and comparison of method performance on additional data not used in context-specific model extraction. Therefore, our study sets the ground for applications of other regularization techniques in large-scale metabolic modeling.

2.1 Introduction

The investigation and understanding of cell metabolism has experienced a paradigm shift, which has been largely propelled by the development of high-throughput methods in the last two decades. As a result, the classical pathway-centered view has been substituted by a network-driven perspective, which considers the entire set of known interconnected biochemical reactions. This had led to the creation of genome-scale metabolic models (GEMs) for organisms from each of the three domains of life: archaea, bacteria and eukarya (Schellenberger, Park, Conrad, & Palsson, 2010). While a GEM constitutes an organized and comprehensive system of knowledge about an organism, it also allows *in silico* analyses based on constraint-based methods, relying on the corresponding stoichiometric matrix representation and assumptions about cellular metabolism. The findings from these analyses provide useful insights in metabolism, and may circumvent the drawbacks of estimating fluxes from labeling studies–still a computationally demanding undertaking (Kruger, Masakapalli, & Ratcliffe, 2012; Ravikirthi, Suthers, & Maranas, 2011; Young, Shastri, Stephanopoulos, & Morgan, 2011). Furthermore, several methods facilitate the integration of high-throughput data in GEMs. The benefits of these methods are twofold: improving the accuracy of flux prediction and providing a scaffold network for analysis of additional experimental data (Blazier & Papin, 2012; D. R. Hyduke, Lewis, & Palsson, 2013).

However, a metabolic network that includes all known biochemical reactions of an organism may not be realistic in a particular cellular scenario, since there is mounting evidence that cells adapt their metabolism to arising conditions, such as: external environment, developmental stage, cell type in multicellular organisms, or even during a pathological condition (e.g., cancer), to name only a few. In these different *contexts*, only a subset of reactions is typically active. Therefore, the shift towards reconstructing context-specific models of cell metabolism has become necessary to provide more accurate and biologically meaningful insights. This is of particular importance when tackling the physiology of multicellular organisms, not only to better understand tissue- or cell-specific metabolism, but as a first step to reconstruct metabolic networks of an entire organism/body, where multiple specialized models are mutually interconnected (Dal'Molin, Quek, Palfreyman, Brumbley, & Nielsen, 2010; Grafahrend-Belau et al., 2013).

Several methods have been proposed to determine context-specific networks, already comprehensively reviewed in Machado & Herrgård, (2014). In general, the methods for extracting a context-specific model from a given GEM integrate high-throughput data from a particular context to select the set of respective active reactions. While these methods differ with respect to their underlying assumptions and mathematical

formulation, they can be classified into three main groups (Robaina Estévez & Nikoloski, 2014), briefly discussed in the following.

GIMME (Becker & Palsson, 2008) and GIM^{3}E (Schmidt et al., 2013) form the first group, whereby first a metabolic functionality (*e.g.,* biomass production) is optimized through Flux Balance Analysis (Orth et al., 2010) (a linear mathematical program), and then the obtained optimal value is employed to constrain a second linear program which aims at minimizing the discrepancies between fluxes and data. The latter is based on selecting a user-defined data-dependent threshold value and then penalizing reactions whose associated data is under the threshold.

The second group comprises iMAT (Shlomi et al., 2008) and INIT (Agren et al., 2012) which use a mixed integer linear program. The binary variables in this formulation select the reaction states (*i.e.*, active or inactive) which are most concordant with the associate data state. While iMAT uses data to pre-classify reactions of the GEM into active or inactive groups, INIT integrates data as a weighting factor for the binary variable. Moreover, INIT includes a set of key metabolites which must exhibit a small positive deviation from the steady state condition. In an extended version, tINIT (Agren et al., 2014), a set of metabolic tasks (*i.e.*, biochemical pathways) that must carry non-zero flux can be added as further constraints.

The third group, composed of MBA (Jerby et al., 2010), mCADRE (Yuliang Wang et al., 2012) and FastCORE (Vlassis, Pacheco, et al., 2014), first defines a core set of reactions, classified as active in a given context according to experimental data, and then finds the minimum set of reactions outside the core required to satisfy the model consistency condition (*i.e.*, all reactions in the model must be able to carry a non-zero flux in at least one of the allowed steady-state distributions). Unlike the methods in the previous two groups, these only extract a context-specific model and do not provide a respective flux distribution.

With respect to another classification criterion, the first group belongs to the so-called biased methods (within the constraint-based analysis) since the achieved solution depends on the definition of a metabolic function to be optimized. In contrast, the second and third group consist of unbiased methods since they are independent of a metabolic function (Lewis, Nagarajan, & Palsson, 2012). However, in the case of iMAT, MBA and FastCORE, a group of preferential reactions (to the context of choice) must still be predefined. The choice of unbiased methods is of particular importance when the metabolic functions to be optimized under a given context may be difficult to obtain and justify, *e.g.,* in multicellular organisms, where cells perform various specialized functions.

However, none of these methods allow fully automated model extraction and flux prediction, *i.e.,* without using *a priori* knowledge of a context-specific function and without any binarization or pre-classification of reactions in the process of data

integration. This is particularly important in settings where: (*i*) a large number of context-specific models are to be extracted, or (*ii*) in case of poorly characterized organisms, for which no information regarding the context-specificity of reactions or metabolic function may not be available. Here we present RegrEx, an approach based on a regularized least squares optimization whim aims at extracting context-specific models and providing flux distributions in an automated and unbiased way.

2.2 Results & Discussion

2.2.1 The RegrEx method

Regularization is commonly applied when modeling (*i.e.*, learning) high-dimensional functions from observations, as a means to reduce model complexity (*i.e.*, the number of variables included in the model) and prevent overfitting to background noise. The latter has been shown to considerably improve prediction performance and model robustness. Several regularization methods have already been proposed, including: the Dantzig (Candes & Tao, 2007), the Ridge (McDonald, 2009) and the Elastic Net (Zou & Hastie, 2005) selectors. However, a particular one, the Least Absolute Shrinkage and Selection Operator, abbreviated as LASSO (Tibshirani, 1994), has become very popular in high-dimensional regression problems with n explanatory variables and m observations, where $n \gg m$. This has been largely due to a better performance of the LASSO selector in feature selection (typically obtaining sparse models with a minimum number of explanatory variables) along with the simplicity of the operator, which facilitates its computation (Hesterberg, Choi, Meier, & Fraley, 2008; Vidaurre, Bielza, & Larrañaga, 2013).

The LASSO optimization problem is given in (2.1), below, whereby a weighted ℓ_1-norm on the coefficients, β, as regularization to an ordinary least squares regression with response vector, y^m, and variables, X^{mxn} , is minimized:

$$\min_{\beta \in \mathbb{R}^n} \tfrac{1}{2} \| y - X\beta \|_2^2 + \lambda \| \beta \|_1 \ . \qquad (2.1)$$

The weighting parameter, λ, is usually determined by cross-validation, which offers an unbiased way (*i.e.*, not user-defined and purely based on the data) to find a best suited value with respect to a measure of performance (*e.g.*, mean squared error or coefficient of determination). Regularization by means of the ℓ_1-norm, as generalizations of LASSO, has been already applied in metabolic modeling; for instance, it has been used to reconstruct biochemical networks from time series data (Pan, Yuan, & Stan, 2012), as an alternative to more computationally expensive methods, to study network adaptation to mutations (Tirthankar Sengupta. Shivi Jain. Mani Bhushan., 2013) and, more recently, in FastCORE, one of the existing algorithms to reconstruct context-specific models (Vlassis, Pacheco, et al., 2014).

The Regularized Context-specific model Extraction method (RegrEx) aims at finding a feasible flux distribution, *v*, *i.e.*, satisfying the mass-balance, thermodynamic and capacity constraints, which is as close as possible to the experimental data, *d, e.g.*, gene expression or protein level profiles. At the same time, it excludes the reactions irrelevant for the given context by shrinking their fluxes to zero. This is obtained by minimizing the squared Euclidean distance (second norm) between *v* and *d*, and exploiting the ability of the ℓ_1-norm regularization to perform feature selection. This leads to the optimization problem in (2.2), which is analogous to the formulation of LASSO in (2.1) augmented by the cellular constraints, and is given by:

$$\begin{aligned} &\min_{v\in\mathbb{R}^n} \tfrac{1}{2} \| d - v \|_2^2 + \lambda \| v \|_1 \\ &s.t. \\ &1.\ Sv = 0 \\ &2.\ v_{irr} \geq 0 \\ &3.\ v_{min} \leq v \leq v_{max} \end{aligned} \quad , \qquad (2.2)$$

where irreversible reactions, v_{irr}, are forced to taken non-negative flux values.

To implement RegrEx in existing mathematical programming solvers, we cast the optimization problem in (2.2) as a quadratic program, (2.3), which minimizes the second norm of the error vector, $\epsilon = d - v$, considering only the subset, R_D of reactions to which data can be associated—via the GPR associations (Jensen, Lutz, & Papin, 2011). We also need to introduce special constraints to deal with reversible reactions, which can take negative values, while the data vector is always non-negative. To this end, reversible reactions are split into the forward and backward directions; the net flux is then given by the difference of the respective fluxes. In addition, we need to include a vector of binary variables, *x*, to select either the forward or the backward direction for a particular reaction. This is included to avoid the drawback of bounding the two irreversible reactions to the same data value, which would cause the net flux to be zero.

$$\begin{aligned} &\min_{\varepsilon, v\in\mathbb{R}^n} \tfrac{1}{2} \| \varepsilon \|_2^2 + \lambda \| v \|_1 \\ &s.t. \\ &1.\ Sv = 0 \\ &2.\ v_i + \varepsilon_i = d_i, \quad i \in R_D \\ &3.\ v_{irr} \geq 0, \\ &4.\ v_{min} \leq v \leq v_{max} \\ &5.\ -\varepsilon_{\max} \leq \varepsilon \leq \varepsilon_{max} \end{aligned} \quad . \qquad (2.3)$$

Altogether, this results in a mixed integer quadratic program (MIQP) capturing the RegrEx method, (2.4), in which the sign of the net flux for the reversible reactions is part of the optimization problem and relies ultimately on maximizing similarity to data.

$$
\begin{aligned}
& v_{opt} = \underset{\substack{\epsilon=[\epsilon_{irr};\epsilon_{for};\epsilon_{rev}]\in\mathbb{R} \\ v=[v_{irr};v_{for};v_{rev}]\in\mathbb{R}_0^+, \\ x\in\{0,1\}^n}}{\arg\min} \tfrac{1}{2}\|\epsilon\|_2^2 + \lambda \| v \|_1 \\
& s.t. \\
& 1.\ S_{ext} v = 0 \\
& \left.\begin{aligned}
& 2.\ v_{irr_i} + \epsilon_{irr} = d_{irr} \\
& 3.\ v_{for_i} + \epsilon_{for} - x_i d_{revRxns} = d_{revRxns} \\
& 4.\ v_{rev_i} + \epsilon_{rev} - x_i d_{revRxns} = 0
\end{aligned}\right\}, \quad i \in R_D. \\
& 5.\ v_{irr\min} \le v_{irr} \le v_{irr\max} \\
& 6.\ v_{for} + x v_{for\min} \ge v_{for\min} \\
& 7.\ v_{rev} - x v_{rev\min} \ge 0 \\
& 8.\ v_{for} + x v_{for\max} \le v_{for\max} \\
& 9.\ v_{rev} - x v_{rev\max} \le 0
\end{aligned}
\tag{2.4}
$$

As pointed out above, cross-validation is the canonical method to determine an optimal λ-value for a regression problem. However, this is not an appropriate method for RegrEx due to its particular characteristics. Specifically, a consecutive sampling of the observations would imply selecting arbitrary subsets of reactions in the GEM, which may be incompatible a steady-state. For this reason, we optimized λ selection by running the algorithm for a sequence of λ-values and taking the one that rendered the highest Pearson product-moment correlation between fluxes and data (Figure 2.1).

2.2.2 Evaluation of RegrEx performance

As a case study, we applied RegrEx to extract 11 context-specific human models, namely, adipose tissue, colon, heart, hypothalamus, kidney, liver, lung, ovary, skeletal muscle, spleen and testes, and to obtain the corresponding flux distributions. The starting GEM was the Recon 1 reconstruction (Duarte et al., 2007), which was reduced to its consistent part[4]. This pre-processing improves RegrEx performance since the existence of blocked reactions in the GEM would lead to stoichiometric inconsistencies. We used RNA-Seq expression profiles for 11 human tissues as context-specific data (Krupp et al., 2012). Although it has been shown that gene expression does not always represent a good proxy of the metabolic flux state (Moxley et al., 2009; Rossell et al., 2006), it still constitutes the best data source regarding coverage and quality (typically providing quantitative data for the great majority of genes in the GEM).

[4] GEMs may contain blocked reactions, which are reactions incapable of carrying a non-zero flux in any feasible steady-state distributions. This is the case, for instance, if a reaction leads to a dead-end metabolite, which is not consumed by any reaction. The consistent part of a GEM is then obtained by removing all blocked reactions and dead-end metabolites.

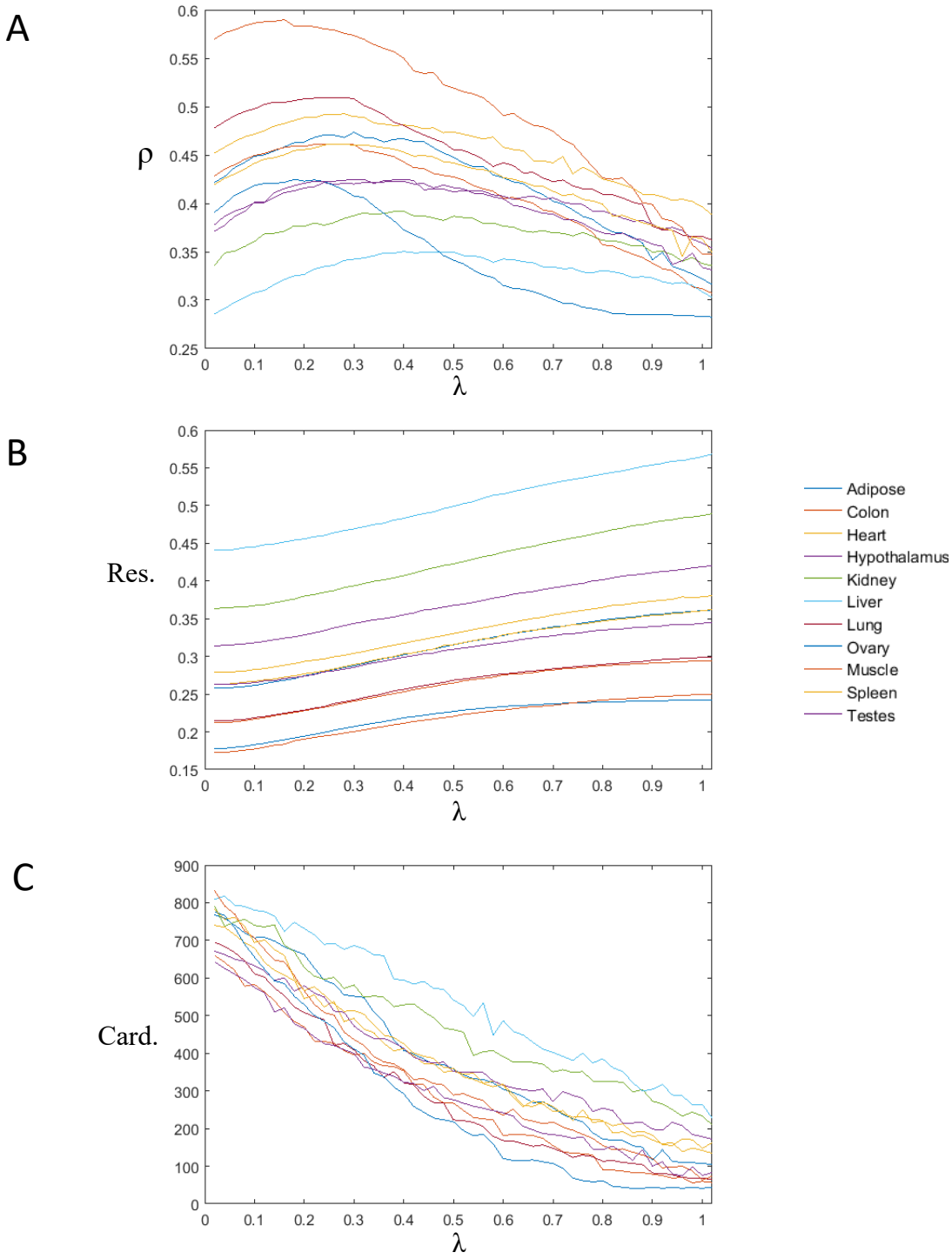


Figure 2.1. The optimal λ value maximizes the correlation between data and flux values. *(A) The Pearson correlation (ρ) is plotted as a function of λ for all human tissues. The correlation increases up to a maximum, at the optimal λ value, then decreases for higher λ values, as the flux through reactions shrinks to zero. (B) In contrast to the Pearson correlation, the residual (Res.), which is computed as averaged absolute difference between data and flux values, increases monotonically with λ. (C) The cardinality of the context-specific models (number of active reactions) decreases with increasing λ, as expected.*

We compared RegrEx performance with other existing methods for context-specific model extraction and flux prediction. Namely, iMAT (Shlomi et al., 2008), the method proposed in Lee et al., (2012)—here called Lee2012—as well as a non-regularized version of RegrEx (RegrEx-λ_0), as a special case of the RegrEx method with $\lambda = 0$. We also included FastCORE in our comparison; however, since FastCORE only provides a context-specific model (*i.e.* set of active reactions), we only included it in comparisons at the structural level of the extracted models. In each case, we used the consistent part of Recon 1 and the same gene expression data from the RNA-Seq Atlas to provide unbiased and fair comparison.

We would like to point out that Lee2012 was not originally developed to extract context-specific models. However, RegrEx has a form similar to that of Lee2012, which aims at improving flux prediction through minimizing the absolute distance between data (*e.g.*, RNA-Seq expression profiles) and flux values. For this reason, we also included Lee2012 in the comparative analysis. Nevertheless, RegrEx differs from Lee2012 in the inclusion of regularization and also in the treatment of reversible reactions: Lee2012 applies an iterative approach, where the optimization problem starts with the subset of irreversible reactions, and reversible reactions are then added sequentially by solving additional optimization problems. This last step is time consuming because it involves two optimization problems per reversible reaction. In contrast, RegrEx selects direction of reversible reactions at once through the use of a binary variable, as explained in Methods (section 2.4), thus reducing the computational time. Moreover, RegrEx is unbiased with respect to the order in which the reversible reactions are added, which is a shortcoming not resolved in Lee2012.

The performance analysis was divided into two parts: First, the similarity with the expression data used to extract the models was evaluated. This evaluation included two measures: the correlation between predicted fluxes and data values (except for FastCORE) and the level of agreement between the correlation matrix of the expression data for each context and the Jaccard distance matrix of the extracted models. (As to quantify the distance between two models in terms of the set of active reactions). Second, we performed an independent validation of the extracted models by measuring their level of agreement with protein expression data taken from the Human Protein Atlas (Uhlén, 2015).

2.2.3 Main characteristics of extracted models by the evaluated methods

The general characteristics of the extracted models by each method are summarized in Table 2.1, and fully detailed in S1 Table. In terms of cardinality (*i.e.*, the number of reactions included in an extracted model), Lee2012 generates models with the lowest mean cardinality (on average approximately 785 reactions per model) followed by RegrEx and RegrEx-λ_0 (with, on average, approximately 843 and 1030 reactions per model, respectively). In contrast, FastCORE and iMAT result in markedly bigger

models for the corresponding contexts (with, on average, approximately 1358 and 1411 reactions per model, respectively). Each set of context-specific models extracted by a particular method has a core set of reactions shared by all contexts. In addition, each context has an exclusive set of reactions (*i.e.*, reactions that are only present in the examined context). In this sense, RegrEx extracted models have the smallest set of shared reactions, with 299 reactions, and the biggest set of total exclusive reactions (*i.e.*, exclusive reactions over all context), with 332 reactions. These two properties demonstrate that the models extracted by RegrEx are in fact more context-specific than the ones extracted by the other methods, which is confirmed by the mean Jaccard similarity between models. The Jaccard similarity is lowest in the case of RegrEx ($\bar{I}_{JRegrEx}$=0.56, with a standard deviation, $\sigma_{IJRegrEx}$=0.04) in support of the previous claim. On the contrary, Lee2012 generates the greatest core set among the methods evaluated, with 509 shared reactions, and the smallest set of total exclusive reactions, amounting to 140 reactions, This, in turn, makes the Lee2012 models to be the least context-specific ($\bar{I}_{JLee2012}$=0.77, with a standard deviation, $\sigma_{ILee2012}$=0.01).

In general, when extracting a context-specific model it is likely that a subset of the reactions in the original (unspecific) GEM is unbounded by data. This can be due to the absence of GPR rules (either because the reaction is not enzyme catalyzed or because the gene-protein association has not been annotated), or simply because experimental data are missing for that reaction. In any case, it is of interest to minimize the number of included reactions without associated data, here called *data-orphan* reactions, since their inclusion results in uncertainty (given the available data). In this manner, simpler models only containing data-orphan reactions that are required to obtain a good overall match with data are preferred. On this line, we evaluated the aforementioned property by computing the *data-orphan* ratio (*i.e.,* the ratio between the number of incorporated reactions with non-associated and that with associated experimental data) of each of the extracted models. RegrEx extracted models show the second lowest mean *data-orphan* ratio across all methods ($O_{RRegrEx}$=0.34, with a standard deviation, σ_{RegrEx}=0.05), only surpassed by Lee2012 ($O_{RLee2012}$=0.28, with a standard deviation, $\sigma_{Lee2012}$=0.03), Notably, this is only valid when a regularized extraction is used, since in the case of RegrEx-λ_0, the *data-orphan* ratio ranks to the second worst position, only surpassed by FastCORE, with mean orphan ratios of 0.47 and 0.50, respectively. The reduced *data-orphan* ratio indicates that RegrEx, although surpassed by Lee2012, is still capable of extracting compact models in which the number of data-orphan (uncertain) reactions is minimized.

Regarding the set of represented reactions of Recon 1 across all contexts, here called total cardinality, RegrEx models collect 1618 unique reactions out of the total of 2469 reactions in Recon 1. Therefore, RegrEx models rank in an intermediate position between Lee2012 models (with 1092 total unique reactions) and iMAT and FastCORE (with 2205 and 2232 total reactions, respectively). The differences in total cardinality as well as mean cardinality per model can be explained by the two main objectives that the evaluated methods take; namely, minimizing the distance between

data and flux values, like in the case of Lee2012 and RegrEx($-\lambda_0$), or including the entirety or a majority of a predefined core set for a particular context, like in FastCORE and iMAT, respectively, which here was the same for both methods (see Methods, section 2.4).

Indeed, this last grouping was reflected when we compared the similarity between models of the same context extracted by the different methods. Results were similar across all contexts, iMAT and FastCORE-derived models shared many reactions, as indicated by the high Jaccard similarity index (mean value across contexts, $\bar{I}_{JiiMAT/FastCORE}$=0.81, with $\sigma_{IiiMAT/FastCORE}$=0.02) and are thus grouped together in the corresponding dendrogram of Figure 2.2. On the other hand, RegrEx, RegrEx-λ_0 and Lee2012 form another cluster, where models extracted by RegrEx and RegrEx-λ_0 are grouped together, and the ones extracted by Lee2012 are closer to RegrEx-λ_0.

	$\overline{Card.}$	$\bar{O}_R$	$\bar{\rho}_{(V,D)}$	$\bar{R}_{(V,D)}$	$\bar{I}_J$	*Shared*	*Total Exclusive*	*Total Card*
RegrEx	842,91(55.14)	0,34(0.05)	0,42(0.07)	0,29(0.08)	0,56(0.04)	299	332	1618
RegrExλ_0	1030,30(76.32)	0,47(0.04)	0,38(0.08)	0,28(0.08)	0,65(0.03)	490	239	1711
Lee2012	784,6 (26.53)	0,28(0.03)	0,13 (0.05)	-	0,77 (0.01)	509	140	1092
FastCORE	1357,9(39.03)	0,50(0.04)	-	-	0,61(0.05)	503	230	2232
iMAT	1411(41.62)	0,42(0.04)	- 0.17(0.03)	-	0,65(0.04)	611	210	2205

Table 2.1 Comparison of models extracted by the four evaluated contending methods: Mean values across contexts. *Global characteristics of the models are derived by applying RegrEx (with automated determination of λ), RegrEx-λ_0 (i.e., RegrEx without regularization), Lee2012, iMAT and FastCORE. The abbreviations stand for the following: $\overline{Card.}$ denotes mean cardinality, $\bar{O}_R$, mean data-orphan ratio, $\bar{\rho}_{(V,D)}$, mean correlation between data and predicted flux values, $\bar{R}_{(V,D)}$, mean residual value between fluxes and data, $\bar{I}_J$, mean Jaccard index to any other context, Shared, number of shared reactions across all contexts, and Total Exclusive represents total number of exclusively context-specific reactions across all contexts. Values in round brackets correspond to the standard deviation.*

2.2.4 Similarity to data evaluation

When inspecting the correlation between data values and predicted fluxes, RegrEx obtained the first position in the ranking (mean correlation, $\bar{\rho}_{RegrEx}$=0.42, $\sigma_{\rho RegrEx}$=0.07), followed by RegrEx-λ_0 and Lee2012 (with a mean correlation of 0.38 and 0.13, respectively). Moreover, iMAT results in the worst mean correlation value of -0.17 (FastCORE does not provide flux values, as commented before, so it is not evaluated with respect to this criterion). However, this difference in correlation can be explained by the different approach followed by iMAT, since in this case the method does not aim at minimizing the distance between data and flux values.

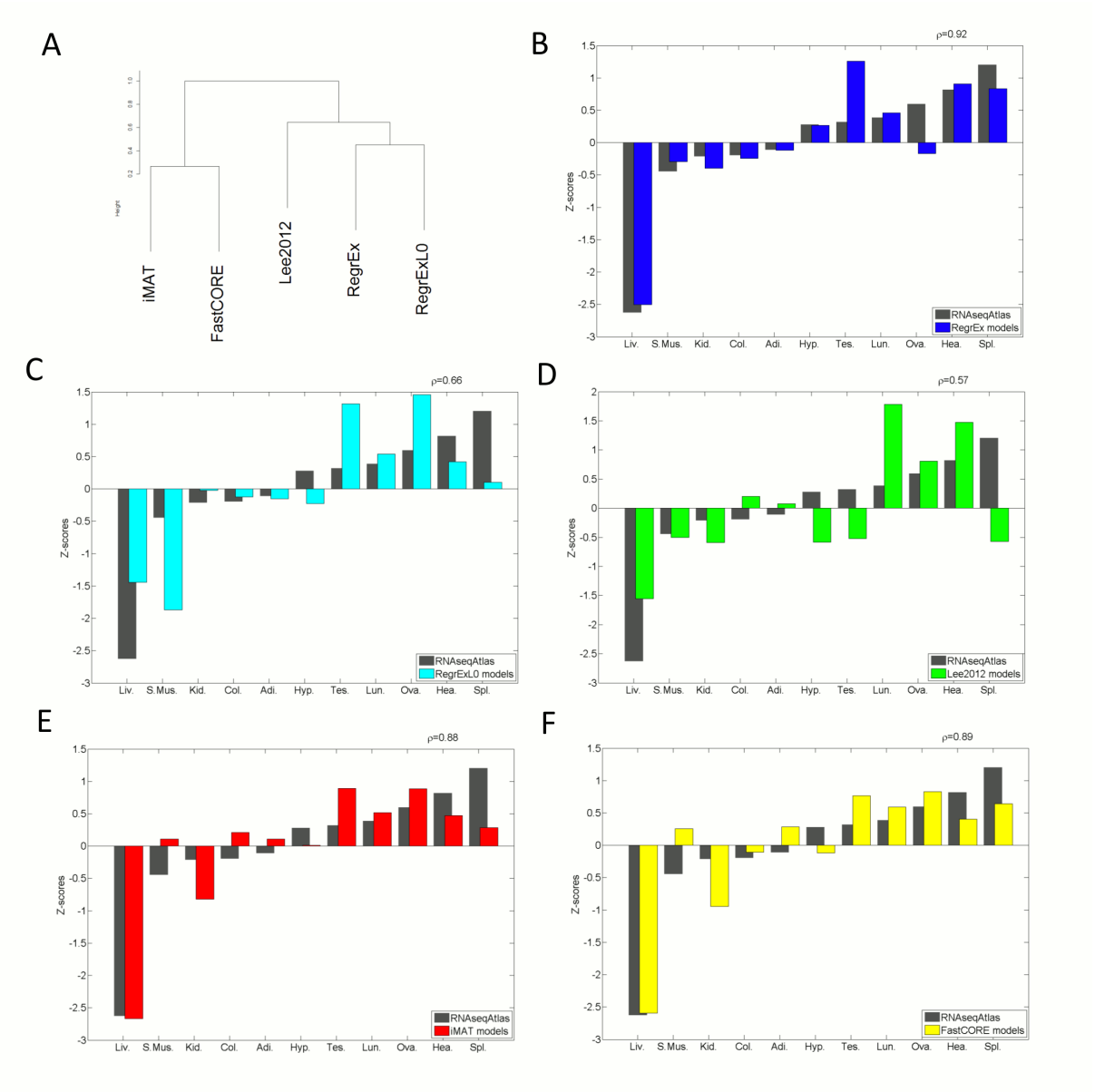


Figure 2.2. Dendrogram clustering the evaluated methods and comparison of data- and model-derived z-scores quantifying the differences between contexts. *(A) The dendrogram is obtained from the the Jaccard similarity that models have across the different methods. Two main clusters are formed, iMAT and FastCORE on one side, and Lee2012, RegrEx-λ_0 and RegrEx on the other. In the second cluster, RegrEx and RegrEx-λ_0 form a subcluster. (B-F) data- and model-derived z-scores are compared for RegrEx, RegrEx-λ_0, Lee2012, iMAT and FastCORE, respectively. Correlation values between the two series (data and model) are shown in the right upper corner in each case. Adi.:Adipose, Col.:Colon, Hea.:Heart, Kid.:Kidney, Liv.:Liver, Lun.:Lung, Ova.:Ovary, S.Mus.:Skeletal Muscle, Spl.:Spleen, Tes.:Testes.*

To include FastCORE in the comparative analysis, we next inspected the similarity to data in a different manner: instead of considering the flux values, we now compared the set of active reactions per context. This criterion captures an aspect of the structure of the extracted metabolic networks. We used the sets of active reactions across different contexts and per method to compute the similarity matrix, using the Jaccard index. We then compared this similarity matrix with the corresponding correlation matrix of the gene expression values for all contexts. Since the compared matrices use different metrics, that is, correlation and Jaccard similarity, we adopted the following procedure for the comparison.Firstly, we computed the column-wise z-scores (as detailed in Methods, section 2.4) for all matrices, both containing correlation values and Jaccard similarities. A high negative z-score corresponds to a context characterized by being highly dissimilar to the rest of the contexts, *i.e.,* showing low correlation or Jaccard similarity values. While a high positive z-score indicates that this context tends to behave in a similar way as the majority of the contexts. The z-scores then summarize the metabolic specificity of the contexts evaluated. Secondly, we computed the correlation between the profile of z-scores derived from the Jaccard similarity matrices and the profile of z-scores of the correlation matrices corresponding to each context and method evaluated. In this manner, we measured how the different methods captured the overall specificity of each context.

RegrEx performed better than iMAT and FastCORE in capturing the pattern showed by gene expression, as quantified by a correlation value of 0.92, between the z-score values of the extracted models and the ones of data, against a value of 0.88 and 0.89 for iMAT and FastCORE, respectively, see Figure 2.2. The better agreement in the case of RegrEx can be observed in the number of mismatches between the sign of the data z-score value and the one of the extracted model. More specifically, RegrEx only fails in the case of the ovary model, which lies under the mean similarity between any two of the extracted models, thus having a negative z-score, while the corresponding expression data lies over the mean. However, iMAT and FastCORE commit three mismatches (skeletal muscle, kidney and colon in FastCORE and skeletal muscle, adipose tissue and hypothalamus in FastCORE). As expected, the performance of RegrEx-λ_0 and Lee2012 is again worse than the other methods, with a correlation value of 0.66 and 0.57, respectively. Interestingly, liver appears to be the most different context in terms of active reactions, and this is captured by all methods except for RegrEx-λ_0. This is not surprising, since liver is well known to be the organ with greater metabolic capabilities.

To conclude this section, it is noteworthy to highlight the comparison of RegrEx performance with the one of its non-regularized version, RegrEx-λ_0, since the only difference between the two approaches is the application of regularization and the inclusion of an optimization step to determine the optimal λ-value. Precisely, the consideration of regularization allows RegrEx not only to extract more compact models, as commented before, but also to increase the correlation between data and

flux values, reduce the data-orphan ratio, and greatly improve the general similarities and differences in the metabolic state of the different contexts.

2.2.5 Evaluation of the models with human protein profiles

We performed an independent test on the biological reliability of the extracted models by all evaluated methods. To this end, we compared the level of agreement of each model with protein expression profiles taken from the Human Protein Atlas (Uhlén, 2015)—we had to exclude hypothalamus from the comparison since it is not present in this database, see Methods, section 2.4. The protein expression data is semi-quantitative, namely, it only provides the categorical levels *high*, *medium* and *low*. To account for this, we evaluated whether models contained an enriched group of genes coding for proteins within the category of *high* expressed for the corresponding organ/tissue in comparison with the other two categories, as well as an enrichment in genes from the *medium* expression value group in comparison to those from the *low* expression.

To test these hypotheses, we determined the number of genes of each group in each context, *i.e.*, the number of genes in *high, medium* and *low* across all organs/tissues, and applied the Mann-Whitney test on the obtained distributions to determine the statistical significance of their difference. This test was applied for each evaluated method. As observed in Table 2.2, RegrEx extracted models are indeed significantly enriched in *high* and *medium* expressed genes since the p-values for all three comparisons, number of genes in the *high* group greater than in the *medium*, H > M, *medium* greater than *low,* M > L, and *high* greater than *low*, H > L, are below the significance threshold of 0.05. In the case of RegrEx-λ_0 and iMAT, only two of the comparisons are significant, H > M and H > L, and M > L and H > L, respectively. Models in FastCORE are only enriched in *high* expressed genes in comparison to *low* expressed and none of the comparisons are significant in the case of Lee2012. These results add an additional experimental support to the extracted models by RegrEx and in a lesser extend the ones by RegrEx-λ_0, iMAT and FastCORE, for two main reasons: the additional experimental data comes from a different (independent) database than the one used during the extraction, and relies on a lower hierarchical level in the causal chain controlling metabolic fluxes, namely, is based on protein expression rather than gene expression.

Method	*H>M*	*M>L*	*H>L*
RegrEx	**0.0445**	**0.0262**	**0.0034**
RegrExL0	**0.0045**	0.1763	**0.0006**
iMAT	0.1399	**0.0444**	**0.0319**
Lee2012	0.3421	0.4559	0.2179
FastCORE	0.0525	0.1575	**0.0216**

Table 2.2 Comparasion on the level of agreement of each extracted model with the Human Protein Database. *P-values for each Mann-Whitney test (with alternative hypothesis H>M, M>L and H>L) are collected here. A significance threshold of 0.05 was used to reject the null hypothesis (p-values<0.05 in bold).*

2.2.6 Functional analysis of RegrEx extracted models

The next step in evaluating RegrEx extracted models was to perform a functional analysis. Concretely, we determined which metabolic functions were important in each context. To this end, we calculated the flux capacity of every reaction in the models, given by the difference between the maximum and minimum corresponding flux value obtained per Flux Variability Analysis (R. Mahadevan & Schilling, 2003). However, here, the only constraints were arbitrary upper and lower bounds for flux values. We stress that the flux capacity value herein defined only quantifies the theoretical range of flux values that a reaction can take in a given network. Hence, the actual flux range of a reaction in a particular metabolic scenario does not have to coincide with the theoretical. However, the flux capacity of a reaction does depend on the topology of an extracted metabolic network. This makes it a reliable proxy to evaluate which reactions are favored in a certain context. In an alternative way, it also allows evaluating which reactions are not influenced by the differences in network topology across contexts, *i.e.*, which are robust in flux capacity irrespective of the different contexts. Note that this set of robust reactions must belong to the core set of shared reactions across contexts.

To facilitate the analysis, we grouped the subsystems into 8 broader metabolic categories: Central metabolism, Amino-acid metabolism, Carbohydrate metabolism, Cofactor and vitamin metabolism, Lipid metabolism, Nucleotide metabolism, Transport and Others, see S2 Table. In addition, we averaged flux capacity values of the reactions in each metabolic subsystem of Recon 1, and in each previously defined metabolic category. In this manner, we obtained the mean flux capacity (MFC) per subsystem or category. We also counted the total number of reactions (TNR) in each category, as an alternative way of quantifying their metabolic importance in the extracted networks.

Marked differences arose when we compared the results obtained by counting the number of reactions per metabolic category against the ones obtained by averaging the flux capacity. For instance, the category with largest number of reactions in all extracted models (using any of the methods) is Transport, similar to the findings in (Thiele et al., 2013). However, if we look at the MFC, Transport, in general, takes a modest position, often surpassed by Central and Carbohydrate metabolism. More specifically, in the case of RegrEx, approximately 43% of the reactions in Liver are assigned to Transport, whereas Transport only contributes with 6.2% to the total MFC of the extracted model. Similarly, Nucleotide metabolism is associated 14% of the total number of reactions while only contributes with 6.5% to the total MFC of Liver. On the contrary, systems with a smaller number of reactions, such as Amino acid and Carbohydrate metabolism (11% and 3.5%, respectively), get a higher contribution to the total MFC (20% and 14%, respectively), see Figure 2.3 and S1 Fig. for a complete comparison for each context.

As commented before, all context-specific, RegrEx extracted models share a core set of 299 reactions. If we consider the TNR in each metabolic category we obtain the distribution in Figure 2.3. The core is dominated by Transport reactions (46%), followed by Nucleotide metabolism (21%) and Central metabolism (17%), being the rest of the categories represented to a smaller extent. Moreover, when computing the MFC for each individual reaction in the core, we see that the majority of them (80%) are robust reactions. Where robust here means that the flux capacity is maintained across contexts. However, a non-negligible part of the core (the remaining 20%) is constituted by non-robust reactions—those that, although being shared by all contexts, present a variable flux capacity. In this group, we encounter reactions like the superoxide dismutase (ROS detoxification), with a coefficient of variation (CV) value of 0.49, one of the highest in the core. Interestingly, we also find all the reactions belonging to the pentose phosphate pathway that are present in the core (see S3 Table). These observations show that not all reactions in the core behave in a similar way. On the contrary, we can partition it into a subset of reactions that are independent of (context-specific) modifications of the network topology, and a subset that depends on the context and therefore can be more or less prominent in certain tissues or organs.

Alternatively, to further evaluate the functional validity of the RegrEx extracted models, we used the previously calculated MFC to investigate the importance that a given subsystem had in each context. Furthermore, we ranked the subsystems according to the CV of the MFC value distribution of each subsystem across contexts. This implies that subsystems with a low CV are evenly represented among the different contexts, while with increasing CV, these subsystems tend to be more specific for certain contexts. For instance, all subsystems belonging to Central metabolism occupy top positions in the ranking, which is expected due to the fundamental role that these subsystems play in all cell types. The citric acid cycle is

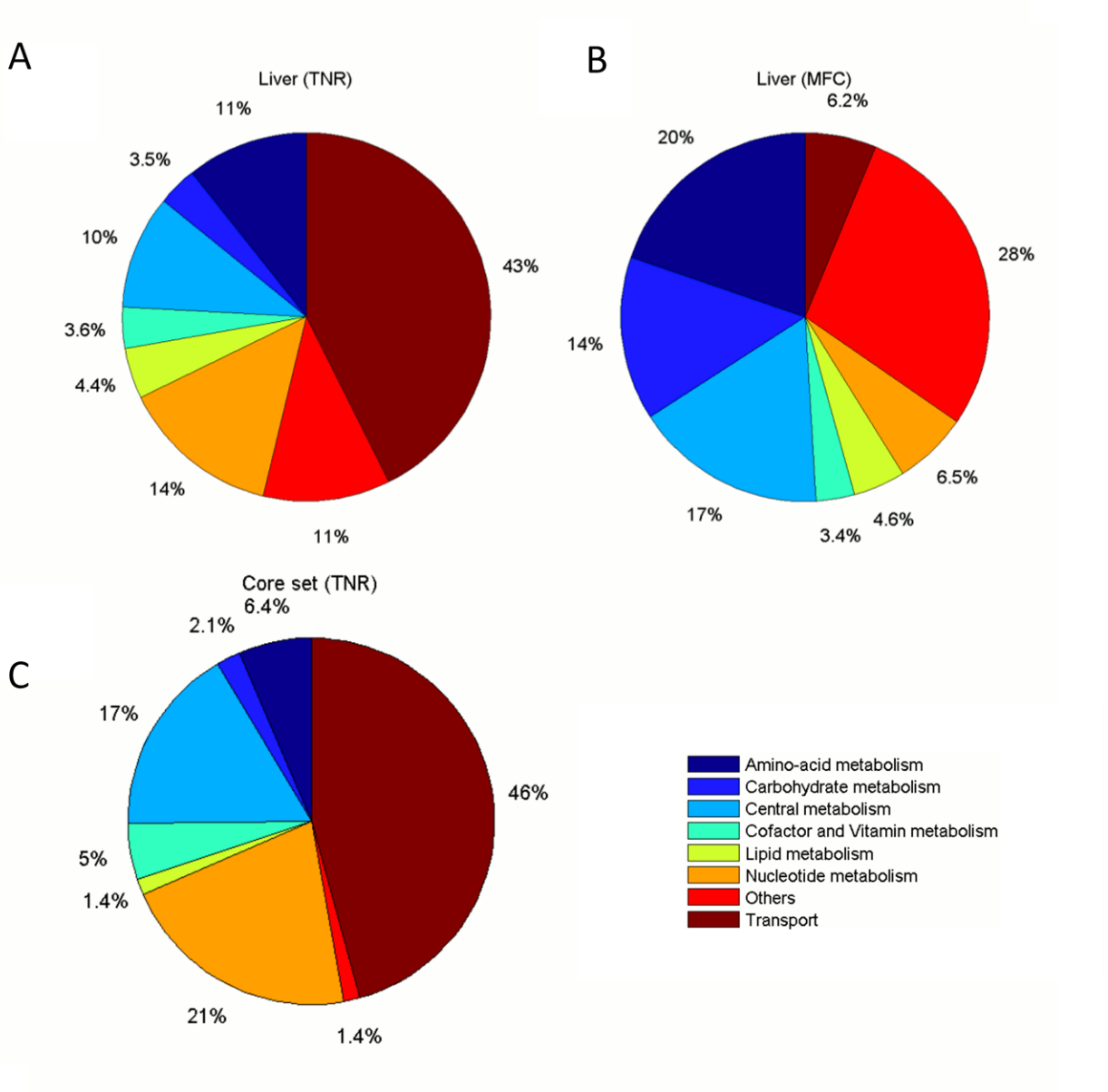


Figure 2.3. Pie charts displaying the core set of shared reactions for the 11 models extracted by RegrEx and a selected comparison of the metabolic categories presented in the liver model. *(A) Distribution of the total number of reactions (TNR) per metabolic category for the liver model (containing a total of 821 reactions across all categories). (B) distribution of the mean flux capacity (MFC) values of each metabolic category in the liver model, as explained in the main text, noticeable differences arise with respect to the distribution depict in A. (C) reaction content (TNR) for the core set of shared reactions divided per metabolic category, as explained in the main text, the three dominant categories are Transport, Central and Nucleotide metabolism. Metabolic category names are displayed in the color bar legend.*

the first subsystem in the ranking with a CV value of 0.078. On the contrary, the pentose phosphate pathway is the subsystem in Central metabolism with a highest CV value of 0.21. However, it can be considered low in the context of the entire ranking,

and may be explained by the fact that, unlike the rest of subsystems in Central metabolism, the totality of its reactions in the core are non-robust, as mentioned before, see Figure 2.4 and S4 Table. In addition to these subsystems, we also find in top positions others equally fundamental pathways, including: NAD, folate and vitamin A metabolism (all in the category of Cofactor and Vitamin metabolism), extracellular and mitochondrial transport or nucleotides metabolism. Interestingly, the last three subsystems also contain the greatest number of the previously defined robust reactions (S4 Table).

Lipid metabolism presents a middle level of specialization across the different contexts, which is reflected by its middle positions in the ranking. Therefore, this finding implies that there are some tissues in which lipid metabolism is predominant. For instance, fatty acid metabolism is predominant in adipose tissue, liver and skeletal muscle. In addition, fatty acid oxidation and fatty acid activation are predominant in adipose tissue and skeletal muscle. This is consistent with known functions of these contexts; the adipose tissue and liver are primary locations for fatty acid metabolism, while the fatty acid oxidation provides the required energy supply for oxidative muscle contraction (Frayn, Arner, & Yki-Järvinen, 2006). We also find gluthatione metabolism (assigned to the category "Others") in a middle position, with a CV of 0.6, and it is highlighted in kidney. This last feature also serves as validation of the extracted model, in fact, glutathione metabolism is essential to the kidney for an adequate functioning (Lash, 2005). Finally, the lasts positions are mainly populated by subsystems in Cofactor and Vitamin metabolism and the miscellaneous category "Others", such as: keratin sulfate degradation, heme biosynthesis and degradation or bile acid biosynthesis, see S4 Table and Figure 2.4

The subsystems with largest CV value include those with extreme behavior, *i.e*., these subsystems are only predicted to be active in a single context. This category consists of: bile acid biosynthesis, biotin, riboflavin, vitamin B6, vitamin D, CYP, methionine and D-alanine metabolism (S4 Table). We can explain this behavior as a reflection of the original gene expression values associated to the reactions in each of these subsystems. For instance, in the case of bile acid biosynthesis, the liver presents an extremal value in the distribution of expression values in the subsystem across contexts (z-score=2.9, S5 Table). We wanted to know if this characteristic was sufficient to explain the artifact, or, in contrast, the network topology of Recon 1 was also contributing to this observation. The latter may happen if some reactions crucial to satisfying the steady-state conditions were missing in Recon 1. To test this hypothesis, we applied RegrEx on Recon 2, a recent extended version of the Recon 1 model of increased size, *i.e.,* 5317 in Recon 2 *versus* 2469 reactions in Recon 1, both after eliminating the blocked reactions (Thiele et al., 2013).

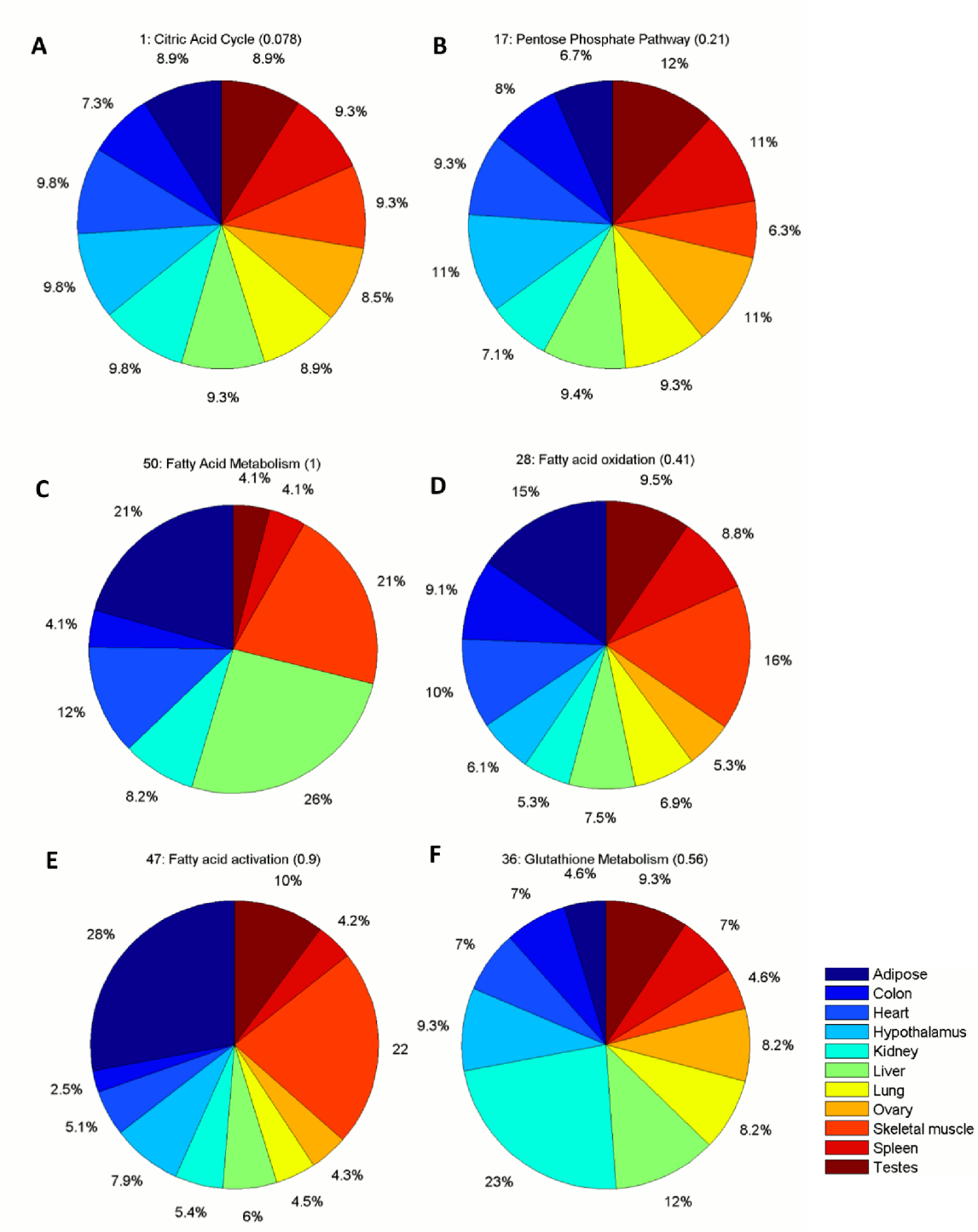


Figure 2.4. Illustration of selected Recon 1 subsystems displayed in a Pie chart form, depicting the distribution of mean flux capacity (MFC) values across contexts. Panels *A-B correspond to the two extreme metabolic subsystems, in terms of CV, in Central metabolism. The citric acid cycle (A) shows the lowest CV value (both in Central metabolism and within the entirety of Recon 1 subsystems). The pentose phosphate pathway (B) shows the greatest CV value in Central metabolism. (C-E) the distribution of MFC values is shown for fatty acid metabolism (C) which is predominantly represented in liver, adipose tissue and skeletal muscle, fatty acid oxidation (D) and fatty acid activation (E) both subsystems predominant in adipose tissue and skeletal muscle. (F) the MFC distribution across contexts is depicted for gluthatione metabolism. Kidney is the context where this subsystem gets a highest MFC value, constituting a 23 % of the total MFC value across contexts. See main text for details. In all cases, the first number preceding the name of the subsystem corresponds to its position in the ranking generated by the CV values, which are shown in round brackets here. Context names are displayed in the color bar legend.*

After extracting the context-specific models, and ranking the subsystems by the CV value of the MFC, we found that the majority of these subsystems were now represented in more than one context. For instance, bile acid biosynthesis (named bile acid synthesis in Recon 2) is present in all contexts, but has a greatest MFC value in Liver, see Figure 2.5 and S6 Table.

2.2.7 Computation time comparison

RegrEx computational performance was also evaluated and compared to the other methods. In all cases, with the exception of FastCORE, the Gurobi (Gurobi Optimization, 2017) solver was used. CPLEX (IBM, n.d.) was used in FastCORE instead of Gurobi for two reasons: CPLEX is the default solver in the code provided in (Vlassis, Pires Pacheco, & Sauter, 2014), and we believe that the differences in computation time between Gurobi and CPLEX are negligible when solving LP problems, as is the case in FastCORE.

Table 2.3 summarizes the differences in computational time and problem formulation for the evaluated methods. Markedly, iMAT shows a very good computation time, with a mean of 0.16 seconds per model extracted. (This result changes dramatically when using GLPK as solver (Makhorin, 2012), where mean network extraction time is typically above one minute). This computation time is comparable to the one obtained by FastCORE. Lee2012 presents a computation time of around 10 minutes per extraction. (This result again changes dramatically when using GLPK as solver, with computation times of several hours per context extracted). When RegrEx is evaluated for a fixed λ-value (that is, RegrEx-λ_0), the computation is fixed to around 60 seconds per model extracted, this is due to the time limitation constraint imposed to the Gurobi solver, as discussed in Methods, section 2.4. When no time limit is imposed RegrEx gets the worst position among the evaluated methods, this may be due to the complexity inherent of solving an MIQP for a big network like the one of Recon 1. However, the sensitivity to time limit analysis suggests that the improvement obtained by increasing the time limit does not compensate the extra time expended, and even could be reduced to 30 seconds with a similar outcome (see Table 2.4 in Methods, section 2.4).

Finally, in the case of RegrEx, the total computation time depends on the number of λ-values evaluated during the optimization step. For instance, in this case the mean computation time per model extraction stays around 15 minutes, since a sequence of 15 λ-values is used in the optimization. This greater computational time required by RegrEx is explained by the necessity of finding an optimal λ-value to control the regularization during the extraction, which is specific to any particular data set and GEM. However, the total computational time spent by RegrEx still remains within a reasonable range, and, as seen in Results, including regularization is fundamental to increase the overall performance.

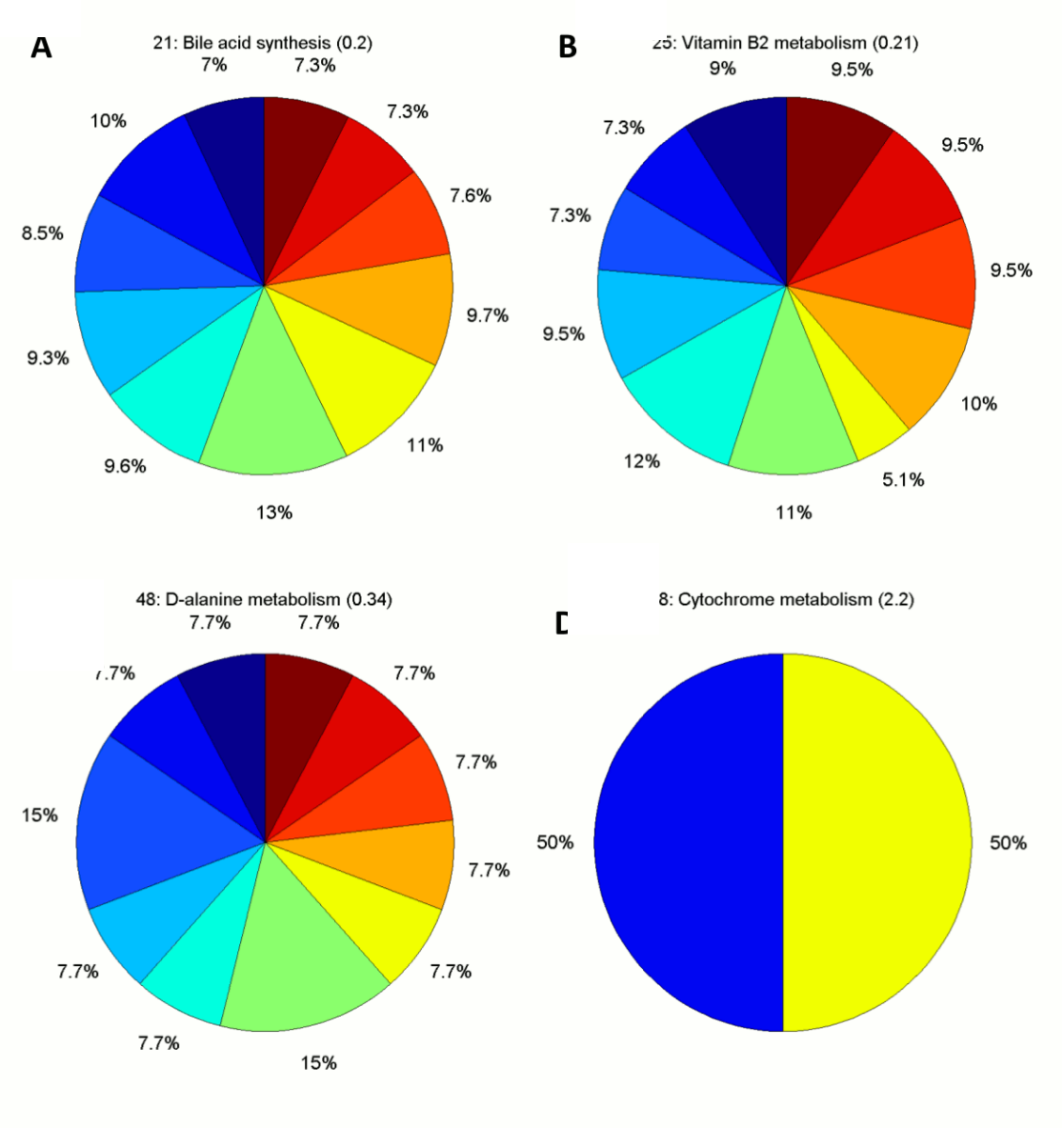


Figure 2.5. MFC value distribution across contexts for selected subsystems in Recon 2. *Bile acid (bio)synthesis (A), vitamin B2 metabolism (B, equivalent to riboflavin metabolism in Recon 1) and D-alanine metabolism (C) are represented in all contexts in Recon 2. Cytochrome metabolism (D, equivalent to CYP in Recon 1) is represented only in two contexts, Lung and Colon, in Recon 2. In all cases, the first number preceding the name of the subsystem corresponds to its position in the ranking generated by the CV values, which are shown in round brackets here. Context names are displayed in the color bar of Figure 2.3. See main text for details.*

Method	*Formulation*	*Solver*	*Mean Time ± SEM **
iMAT	MILP	Gurobi	0.1652 ± 0.0038
FastCORE	LP ↻	CPLEX	0.2976 ± 0.0101
RegrEx-λ0	MIQP	Gurobi	60.0785 ± 0.0025
Lee2012	LP ↻	Gurobi	571.2108 ± 24.8921
RegrEx	MIQP ↻	Gurobi	928.5313 ± 2.3079

Table 2.3 Computation time of the evaluated methods. *Mean computation times per model extraction, type of mathematical program solved and the used commercial solver are displayed for each evaluated method. ↻ stands for iteratively repeated. *Time is shown in seconds, SEM stands for Standard Error of the Mean*

2.3 Conclusion

We have presented RegrEx, a method to extract context-specific metabolic models and provide a flux distribution most in accordance with experimental data. RegrEx generated context-specific flux distributions with the highest correlation values among the competing methods evaluated, as well as extracted compact models, enriched in reactions with high associated data values. Importantly, RegrEx performance is severely impaired when performing a non-regularized extraction (*i.e.,* when $\lambda=0$, here called RegrEx-λ_0). More specifically, the models obtained *without* employing regularization are less specific to each particular context, share a greater amount of reactions and contain less exclusive reactions in comparison to models for other contexts. This is supported by the higher mean Jaccard index over all pairs of compared context-specific models. In addition, the mean orphan ratio is higher if regularization is not used, implying that a greater number of reactions with non-associated experimental data is included and causing these models to be less compact. Finally, the mean correlation values between predicted fluxes and data are also smaller in the non-regularized extraction. Altogether, these observations support the importance of including regularization to obtain a better performance in context-specific model extraction.

RegrEx have also proven to be a suitable method among the alternatives evaluated here, to provide a larger correlation between predicted fluxes and experimental data, as well as models that capture the general pattern of differences and similarities in reaction activity across contexts expressed by data. The models extracted by RegrEx are also in agreement with an independent data source, based on protein expression, and include preferentially genes that are associated to highly expressed proteins, outperforming the competing methods with respect to this criterion.

In the case study presented here, we have used gene expression profiles as experimental data. However, RegrEx can support other data sources; protein profiles can be easily integrated (*e.g.* generated through mass-spectrometry based approaches), and the problem of lower coverage typically presented by protein profiles can be alleviated by jointly integrating gene expression data to fill the gaps. In addition, if there exist strong experimental evidence supporting the presence of a certain reaction in a given context, its lower bound can be set to an arbitrary positive value (*i.e.,* $V_{min} > \epsilon$, when splitting reversible reactions) thus forcing it to be included in the context-specific model. In a similar manner, when the evidence is for the presence of a metabolite, the sum of the reactions producing such metabolite could be constrained to ensure its inclusion, thus allowing integrating metabolomics data in a qualitative way.

RegrEx can be easily used in MATLAB through the provided files. Moreover, no parameters need to be chosen by the user, since the only parameter, λ, is determined by RegrEx in an automated fashion. In this manner, the user only needs to provide a relevant (context-specific) data source(s) and the GEM from where the context-specific model is to be extracted; the rest of the operating process is fully automated. Finally, RegrEx does not require any *a priori* knowledge on metabolic functionality in a given context. The property of being an unbiased method along with the fully automation of the process may be a prominent quality when dealing with complex, multicellular organisms, where multiple cell types or tissues coexist and specialized in certain functions that are not yet very well understood.

2.4 Methods

2.4.1 RegrEx implementation

We solved the MIQP of RegrEx using the Gurobi solver (Gurobi Optimization, 2017). To speed up the optimization, we restricted the computation time to 60 seconds per MIQP. Additional robustness analyses indicated that higher computation times implied a low increase in performance (tripling the time limit, *i.e.,* 180 seconds, only caused a mean correlation increment between models of 0.0004, see Table 2.4).

	30s	*60s*	*90s*	*180s*
$\bar{\rho}_{(V,D)}$	0.4493(.068)	0.4493(.068)	0.4493(.068)	0.4497(.068)
$\bar{R}_{(V,D)}$	0.2845(.081)	0.2845(.081)	0.2845(.081)	0.2844(.081)
$\overline{Card.}$	856.55(92.124)	856.27(91.153)	855.45(91.35)	856.55(93.17)

Table 2.4 Results comparison for different time limits applied to the Gurobi solver*. Four different time limits were evaluated to test the sensitivity of optimal solutions to the early termination criterion (60 s) imposed. In all cases, the λ-value was fixed to a reference optimum, the one obtained when the*

time limit was 60 s. Mean values for the 11 contexts (with the standard deviation within round brackets) are shown for the correlation between flux values and data, $\overline{\boldsymbol{\rho}}_{(V,D)}$*, the mean residual,* $\overline{\boldsymbol{R}}_{(V,D)}$*, and the cardinality,* i.e.*, number of reactions of the extracted models,* $\overline{\boldsymbol{Card}}$*.*

RegrEx was implemented in MATLAB, and the code is provided in the S1 File of the supporting information in (Robaina Estévez & Nikoloski, 2015). The implementation provides the final context-specific models in a COBRA toolbox compatible format (D. Hyduke et al., 2011), thus allowing facile subsequent analysis.

2.4.2 Context-specific model extraction

To test RegrEx performance we selected, as a case study, the existent human metabolic network reconstructions, Recon 1 (Duarte et al., 2007) which has been previously used with other algorithms (Agren et al., 2012; Vlassis, Pacheco, et al., 2014; Zur, Ruppin, & Shlomi, 2010) and Recon 2 (Thiele et al., 2013) as a further test on a larger network. This case study allows a direct comparison between the models extracted by different methods. As input data, we used available RNA-Seq human expression profiles for 11 different contexts (*i.e.*, organs or tissues) published online in the RNA-Seq Atlas (Krupp et al., 2012), and normalized to RPKM values. To avoid blocked reactions, we first extracted the consistent part of Recon 1 through a standard flux variability analysis, using the *reduceModel* function of the COBRA toolbox (D. Hyduke et al., 2011).

The range of expression values typically varies between genes, especially in RNA-Seq-derived expression data, where differences in mean values (*e.g.*, across tissues) between genes can be of several orders of magnitude. This may likely cause RegrEx to favor reactions whose associated genes have higher mean values across contexts, thus, reconstructing context-specific models in a biased manner. To correct for this bias, we normalized the expression value, t, of each gene, i, in context, j, to its maximum value across all considered contexts:

$$d_{i,j} = \frac{t_{i,j}}{\max(t_{i,\forall j})}, \qquad \mathrm{i} \in genes, j \in contexts. \tag{2.5}$$

2.4.3 Performance analysis with competing methods

The existing iMAT implementation in the COBRA toolbox (D. Hyduke et al., 2011) was used to perform the iMAT model extraction. The 75th percentile of the cumulative distribution was used as a threshold to binarize the gene expression data, *i.e.*, to create the high- and low-expressed (reaction associated gene/s) groups. The implementation provided in (D. Lee et al., 2012) was used to analyze the model extraction approach, denoted as Lee2012. Since the RNA-Seq Atlas does not provide any variance measurement, the weighting factor to correct for experimental error was

not used. In addition, the upper bound on flux values was set to 1, as in RegrEx, for fair comparison. Reactions with an absolute value above 10^{-6} were considered active for a given context. In the case of FastCORE, we used the implementation provided in (Vlassis, Pires Pacheco, et al., 2014) and obtained the core set of reaction by taking the reactions with an expression value for the associated gene(s) above the 75^{th} percentile; therefore, it uses the same set as the high-expressed group employed in iMAT.

The Jaccard index was used to generate the similarity matrices comparing models extracted for different contexts, as well as models of the same context extracted by different methods. In this last case, the clustering dendrogram in Figure 2.2 was generated with the *hclust* function of the package *stats* in the R environment, and by using the average linkage criterion

We z-normalized the sum of Jaccard similarities of each context to the remaining (*i.e.*, the sum of each column of the distance matrix). Therefore, the respective z-score quantifies the extent to which a given context differs from the remaining.

2.4.4 Model agreement with human protein expression data

The protein expression profiles were taken from the Human Protein Atlas (Uhlén, 2015) where 10 out of the 11 contexts are represented (note that the hypothalamus is missing, so we did not include it in the evaluation; moreover, for the adipose tissue, we took the data from the cell type adipocyte). In the Human Protein atlas, protein expression levels are classified into *high*, *medium* or *low* and are derived by inmunohystochemical staining. Recon 1 uses Entrez gene identifiers, while the protein coding genes in the Human Protein Atlas are identified following the Ensembl convention. Hence, we mapped the Ensembl identifiers onto Recon 1 using the BioMart data mining tool from Ensembl ("BioMart (Ensembl)," n.d.).

Acknowledgments

We thank the anonymous reviewers for providing an excellent and constructive critique that has improved the quality of the presented work.

Contributions

Performed the experiments: SRE. Analyzed the data: SRE, Wrote the paper: SRE, ZN. Developed the method and wrote the code: SRE. Proposed the original idea and helped during the development: ZN

Chapter 3

Alternative optima in context-specific metabolic predictions

Published as:

On the effects of alternative optima in context-specific metabolic model predictions

Semidán Robaina Estévez[1,2], Zoran Nikoloski[1,2]

[1]Systems Biology and Mathematical Modeling Group, Max Planck Institute of Molecular Plant Physiology, Potsdam-Golm, Germany and [2]Bioinformatics group, Institute of Biochemistry and Biology, University of Potsdam, Potsdam-Golm.

PLoS Comp. Biol. (2017), DOI: 10.1371/journal.pcbi.1005568

Abstract

The integration of experimental data into genome-scale metabolic models can greatly improve flux predictions. This is achieved by restricting predictions to a more realistic context-specific domain, like a particular cell or tissue type. Several computational approaches to integrate data have been proposed—generally obtaining context-specific (sub)models or flux distributions. However, these approaches may lead to a multitude of equally valid but potentially different models or flux distributions, due to possible alternative optima in the underlying optimization problems. Although this issue introduces ambiguity in context-specific predictions, it has not been generally recognized, especially in the case of model reconstructions. In this study, we analyze the impact of alternative optima in four state-of-the-art context-specific data integration approaches, providing both flux distributions and/or metabolic models. To this end, we present three computational methods and apply them to two particular case studies: leaf-specific predictions from the integration of gene expression data in a metabolic model of *Arabidopsis thaliana*, and liver-specific reconstructions derived from a human model with various experimental data sources. The application of these methods allows us to obtain the following results: (*i*) we sample the space of alternative flux distributions in the leaf- and the liver-specific case and quantify the ambiguity of the predictions. In addition, we show how the inclusion of ℓ_1-regularization during data integration reduces the ambiguity in both cases. (*ii*) We generate sets of alternative leaf- and liver-specific models that are optimal to each one of the evaluated model reconstruction approaches. We demonstrate that alternative models of the same context contain a marked fraction of disparate reactions. Further, we show that a careful balance between model sparsity and metabolic functionality helps in reducing the discrepancies between alternative models. Finally, our findings indicate that alternative optima must be taken into account for rendering the context-specific metabolic model predictions less ambiguous.

3.1 Introduction

Genome-scale metabolic models (GEMs) have proven instrumental in characterizing the activity of metabolic pathways in different biological scenarios. The activity of all metabolic reactions is specified by the flux distribution, which can be readily inferred from GEMs through the usage of constraint-based approaches (Bordbar, Monk, King, & Palsson, 2014; Lewis et al., 2012). Such approaches often infer fluxes as solutions to a convex optimization problem in which an objective function is optimized under specified constraints. Two types of constraints can generally be considered: The first is due to the stoichiometry, thermodynamic viability (*i.e.*, if a reaction is irreversible or reversible under normal physiological conditions) and mass-balance conditions. These constraints are included in every constraint-based approach. The second type comprises constraints specific to each approach, and usually reflects the context-specific knowledge or data to be integrated. Flux distributions which satisfy the set of constraints are called feasible. A convex optimization problem is guaranteed to render a unique optimal value (Boyd & Vandenberghe, 2010). However, it is not always guaranteed that there is a unique flux distribution realizing the optimal objective value, which leads to alternative optimal flux distributions. Indeed, such a space of alternative optima arises even in the case of flux balance analysis (FBA), as a classical representative of constraint-based approaches (Binns, de Atauri, Vlysidis, Cascante, & Theodoropoulos, 2015; Kelk et al., 2012; S. Lee et al., 2000; R. Mahadevan & Schilling, 2003; Müller & Bockmayr, 2014; Reed & Palsson, 2004).

Experimental systems biology studies have generated comprehensive atlases of transcript, protein, and metabolite levels from different contexts, such as: cell types, developmental stages, and environments, across different species from all kingdoms of life (Barrett et al., 2013; Kopka et al., 2005; Marx, 2014; Petryszak et al., 2014; Uhlen et al., 2016; Wishart et al., 2013). Analyses of these data sets have already pointed that context-specific differences in the levels of molecular components often affect the activity of metabolic pathways. Additionally GEMs allow constraint-based approaches to integrate such data sets through the so-called gene-protein-reaction rules, which relate metabolic reactions with the enzymes involved and their coding genes (Blazier & Papin, 2012; Machado & Herrgård, 2014; Maria P. Pacheco, Pfau, & Sauter, 2016; Robaina Estévez & Nikoloski, 2014). These approaches address two aims: (*i*) obtaining context-specific flux distributions and (*ii*) determining context-specific GEMs; we refer to the respective approaches as flux- and network-centered, respectively. Alternative optima may also result from the integration of context-specific data. In both settings, the existence of alternative optima leads to ambiguity in context-specific flux distributions and/or network reconstructions, since alternative solutions may substantially differ. This is particularly important in the case of context-specific network reconstructions, where further investigations (e.g. through

constraint-based approaches) conducted on a single optimal network could lead to erroneous conclusions.

To our knowledge, only three studies considered the space of alternative optimal solutions arising from flux-centered approaches: The approach termed iMAT (Shlomi et al., 2008) proposed a procedure to classify the flux state of reactions into active, inactive or uncertain across the alternative optima space. Another approach, abbreviated as EXAMO (Rossell et al., 2013), later used the set of active reactions obtained from the iMAT alternative optima space as input to the approach referred to as MBA (Jerby et al., 2010), a network-centered method, to reconstruct a context-specific network. Additionally, the Flux Variability Sampling (Recht et al., 2014) was used to sample the alternative space of flux values that are equidistant to the data integrated. Finally, we note that alternative optimal context-specific models have not been recognized in the case of network-centered approaches, and currently, there is no available method for their analysis.

In the present study, we propose a method to quantify the variability of alternative optimal flux values of a flux-centered approach. Additionally, we quantify the effect in the alternative optima of including an additional constraint in the flux values, minimize the total sum of absolute flux values, which has been proposed to obtain unique solutions in a flux-centered method (Collins, Reznik, & Segrè, 2012). Furthermore, we investigate, for the first time, the space of alternative optimal context-specific models that arise from several network-centered approaches, and analyze the potential impact on further metabolic predictions and drawn biological conclusions. The study is organized in two parts. The first part is dedicated to explaining the mathematical and computational logic of both (*i*) the context-specific data integration approaches herein evaluated, and (*ii*) the methods that we propose to analyze the respective alternative optima. The second part presents the findings obtained from applying the previously described methods to two particular case studies: a leaf-specific reconstruction from the model plant *Arabidopsis thaliana*, and a human liver reconstruction. This second part serves as an illustration of the impact that alternative optima have in context-specific metabolic reconstructions, and may be followed independently from the first part—which is primary addressed to the specialized reader.

3.2 Results and discussion

3.2.1 Evaluation of alternative optima: Computational methods

In this section, we present the mathematical formulation of the computational methods that we developed to investigate the alternative optima of three selected data integration approaches. In all three cases, we first provide an overview of the approach, which is followed by a description of the method to explore its alternative optima space. We start by a representative of a flux-centered approach—a modified

version of RegrEx (Robaina Estévez & Nikoloski, 2015)—and the method that we propose to explore its alternative optima, termed RegrEx Alternative Optima Sampling ($\text{RegrEx}_{\text{AOS}}$). We then focus on Core Expansion (CorEx), also developed in this study, which we take as representative of a network-centered approach. In addition, we show how the optimization problem behind CorEx can be adapted to evaluate not only its alternative optima space, but that of FastCORE (Vlassis, Pacheco, et al., 2014) and CORDA (Schultz & Qutub, 2016), two state-of-the-art network-centered approaches.

3.2.1.1 Alternative optima in flux-centered approaches: the case of RegrEx

Background

Given a GEM and (context-specific) gene or protein expression data, the Regularized metabolic model Extraction (RegrEx) method reconstructs a context-specific metabolic model, along with the corresponding flux distribution. To this end, RegrEx finds a feasible flux distribution that is closest to a given experimental data set. Therefore, it can be considered a flux-centered approach.

The original RegrEx approach relied on a regularized least squares optimization, in which the Euclidean distance between the given gene expression data vector, d, and a feasible flux distribution, v, *i.e.*, the squared ℓ_2 norm of the difference vector $\epsilon = d - v$, was minimized (Chapter 2, optimization problem (2.4)). The regularization was implemented by also considering the (weighted) ℓ_1 norm of v in the minimization problem, as a means to select the reactions in the GEM that are most important for a given metabolic context. However, here we used a slightly modified version of RegrEx: Instead of minimizing the sum of square errors, we minimized the sum of absolute errors, *i.e.*, the ℓ_1 norm of ϵ. Except for this substitution, the modified RegrEx version, called $\text{RegrEx}_{\text{LAD}}$ (for Least Absolute Deviations), follows the same formulation as the original RegrEx. This modification is required to guarantee that the optimization problem employed to explore the alternative optima space of RegrEx remains convex. The details of this modification and a comparison with the original RegrEx formulation are provided in Appendix S3.1.

The minimization problem behind $\text{RegrEx}_{\text{LAD}}$ considers a set of constraints required to handle reversible reactions: In this case, absolute flux values must be considered when minimizing the distance to the (non-negative) associated gene expression (*i.e.,* for a reversible reaction i, $\epsilon_i = |v_i| - d_i$). This is accomplished by splitting reversible reactions into the forward and backward directions, each constrained to have non-negative flux value, and introducing a vector of binary variables, x, to select only one of them during the optimization. Altogether, these particularities are captured in the mixed integer linear program (MILP),

$$
\begin{aligned}
& v_{opt} = \arg\min \mathrm{w}^T (\epsilon^+ + \epsilon^-) + \lambda \| v \|_1 \\
& \epsilon^+ = [\epsilon^+_{irr}; \epsilon^+_{for}; \epsilon^+_{back}], \\
& \epsilon^- = [\epsilon^-_{irr}; \epsilon^-_{for}; \epsilon^-_{back}], \\
& v = [v_{irr}; v_{for}; v_{back}] \in \mathbb{R}^+_0, \\
& x \in \{0,1\}^n \\
& s.t. \\
& 1.\ S_{ext} v = 0 \\
& \left.\begin{aligned}
& 2.\ v_{irr_i} + (\epsilon^+_{irr} - \epsilon^-_{irr}) = d_{irr} \\
& 3.\ v_{for_i} + (\epsilon^+_{for} - \epsilon^-_{for}) + x d_{revRxns} = d_{revRxn} \\
& 4.\ v_{rev_i} + (\epsilon^+_{back} - \epsilon^-_{back}) - x d_{revRxn} = 0
\end{aligned}\right\}, \quad i \in R_D. \\
& 5.\ v_{irr\min} \le v_{irr} \le v_{irr\max} \\
& 6.\ v_{for} + x v_{for\min} \ge v_{for\min} \\
& 7.\ v_{back} - x v_{rev\min} \ge 0 \\
& 8.\ v_{for} + x v_{for\max} \le v_{for\max} \\
& 9.\ v_{back} - x v_{rev\max} \le 0
\end{aligned}
\tag{3.1}
$$

In (3.1), the flux distribution, v, is partitioned into the sets of irreversible (v_{irr}), and reversible reactions proceeding into the forward (v_{for}) and backward directions (v_{back}), and the (reaction) columns of the stoichiometric matrix, S_{ext}, are ordered to match the partition of v. In addition, the components of the error vector, $\epsilon_i = \epsilon^+_i - \epsilon^-_i$, ϵ^+_i, $\epsilon^-_i \ge 0$, are split into two non-negative variables, ϵ^+_i, ϵ^-_i, as a way to computationally treat the otherwise required absolute values of the components of ϵ. Thus, the ℓ_1 norm $\|\epsilon\|_1 = \Sigma_i |\epsilon_i|$ is replaced by $\epsilon^+_i + \epsilon^-_i$ in the objective function (ϵ is defined only over the set of reactions with associated data, R_D in (3.1)). Finally, the λ parameter corresponds to the weight of the ℓ_1 norm in the objective function, and is chosen during the optimization as to maximize the Pearson correlation between data and flux values (Chapter 2, section 2.2.1).

The convexity of the optimization problem in (3.1) guarantees finding the minimum distance between experimental data and a feasible flux distribution that is allowed by the constraints. However, it does not guarantee that the resulting flux distribution is the only feasible one that is optimal with respect to a particular context-specific data. This variability in optimal flux distributions may be attributed to two factors. On the one hand, as mentioned above, not all reactions in a GEM are typically associated to data. In contrast to *data-bounded* reactions, there is a set of *data-orphan* reactions comprising non-enzymatically catalyzed reactions, reactions without gene-protein annotation or without associated data for a particular context. Data-orphan reactions do not contribute to the error norm in the RegrEx$_{LAD}$ objective function, described in (3.1), and their flux value can vary as long as v satisfies the imposed constraints and its ℓ_1 norm is preserved. This situation is depicted in Figure 3.1, where the search for a flux distribution v that is closest to the data vector, d, is carried out in the projection

of the flux cone, $K = \{v: Sv = 0, v_{min} \leq v \leq v_{max}\}$, where d resides. On the other hand, the geometry of K may preclude certain reactions to obtain an exact match with the

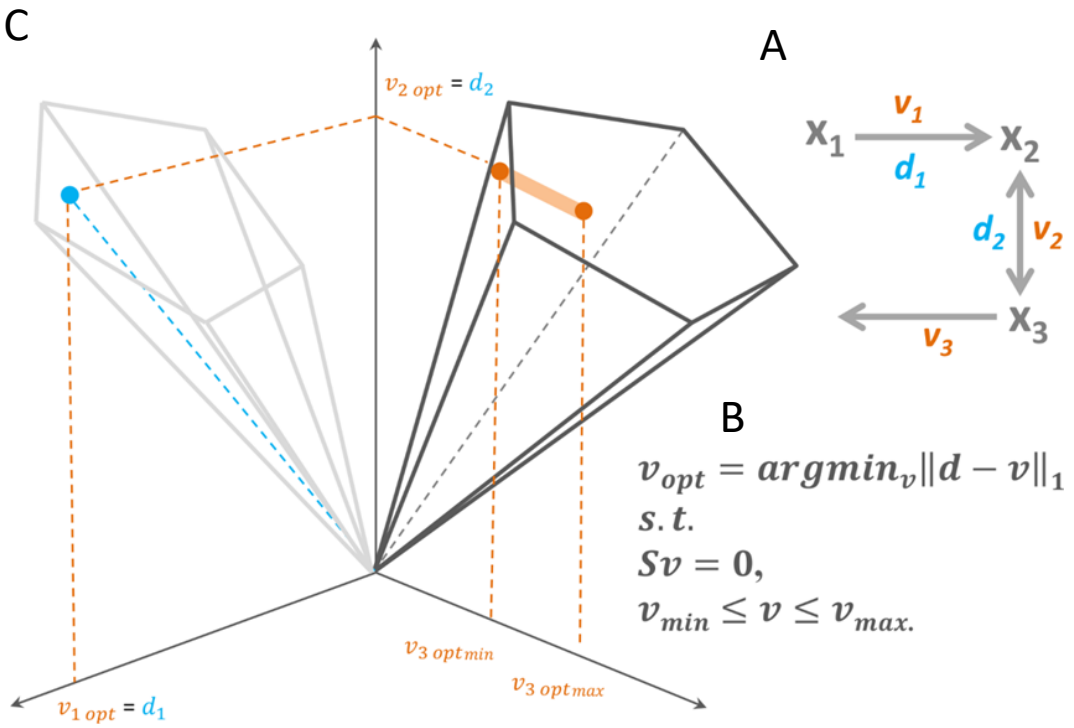


Figure 3.1. A depiction of the alternative optima space of a toy RegrEx data integration problem. *(A) A toy data integration problem for a metabolic network with three reactions, $v_{1\text{-}3}$, and two reaction-associated data values, $d_{1\text{-}2}$ is presented. In RegrEx, the optimization problem consists of finding a flux distribution, v_{opt}, which minimizes the distance to the data being integrated and is compatible with the mass balance and thermodynamic constraints. In this example, only two of the three reactions are data-bounded; thus, the third, v_3, is free to vary its flux value without affecting the minimum overall distance in (B). This situation is depicted in (C), where the flux cone (the set of flux distributions, v, that are compatible with the imposed constraints) is projected onto the two-dimensional space where the data vector, d, resides, and the search for the optimal, v_{opt} is conducted on this projection. This implies that v_3 can vary along the direction orthogonal to the projection plane, as long as its value remains within the flux cone (here depicted as the orange line crossing the cone). Hence, the alternative optima space of this data integration problem consists of alternative vectors, $v_{opt(i)}$, in which the components v_1 and v_2 are fixed, and v_3 varies between $v_{3optmin}$ and $v_{3optmax}$.*

data value, when d remains outside the projection of K. In this case, a set of flux distributions may be equidistant to d, thus generating variability also in the optimal flux value of data-bounded reactions.

The RegrEx alternative optima sampling method

The general approach followed by RegrEx$_{AOS}$, depicted in Figure 3.2, is similar to the Flux Variability Sampling (Recht et al., 2014) (see Appendix S3.1). RegrEx$_{AOS}$ first

creates a random flux vector, v_{rand}, which is bounded by the maximum and minimum flux values previously calculated by Flux Variability Analysis (using only upper and lower bounds as constraints, see Methods, section 3.4). It then searches for the closest flux vector, v, to v_{rand} that belongs to the alternative optima space, *i.e.*, it is at the same distance to the data vector, d, and has the same ℓ_1 norm as the previously calculated $\text{RegrEx}_{\text{LAD}}$ optimum. This is performed by solving the MILP given in (3.2). Finally, $\text{RegrEx}_{\text{AOS}}$ iterates this routine n times to obtain a sufficiently large sample; here we used $n = 2000$, which is sufficient sample size for the subsequent statistical analyses.

$$
\begin{aligned}
&\min_{\substack{\epsilon^+=[\epsilon^+_{irr};\epsilon^+_{for};\epsilon^+_{back}],\\ \epsilon^-=[\epsilon^-_{irr};\epsilon^-_{for};\epsilon^-_{back}],\\ \delta^+=[\delta^+_{irr};\delta^+_{for}],\\ \delta^-=[\delta^-_{irr};\delta^-_{for}],\\ v=[v_{irr};v_{for};v_{back}]\in\mathbb{R}^+_0,\\ x\in\{0,1\}^n}} \quad \| \delta^+ + \delta^- + \delta_{back} \|_1 \\
&s.t. \\
&1-9 \quad (3.1) \\
&10.\ \epsilon^+ + \epsilon^- = \epsilon^+_{opt} + \epsilon^-_{opt} \\
&11.\ \| v \|_1 = \| v_{opt} \|_1 \\
&12.\ v_{irr} - (\delta^+_{irr} - \delta^-_{irr}) = v_{rand(irr)} \\
&13.\ v_{for} - (\delta^+_{for} - \delta^-_{for}) - xv_{rand(revRxn)} = 0 \\
&14.\ -v_{back} + \delta_{back} + xv_{rand(revRxn)} = v_{rand(revRxn)}.
\end{aligned}
\tag{3.2}
$$

The optimization problem in (3.2) inherits constraints 1-9 from (3.1) and incorporates two sets of new constraints. Constraints 10 and 11 are added to guarantee that v renders the same similarity to data and the same ℓ_1 norm of the previously found $\text{RegrEx}_{\text{LAD}}$ optimum, v_{opt}, respectively. In addition, constraints 12-14 introduce the auxiliary variables δ_{irr}, δ_{for} and δ_{back} quantifying the distance of an optimal flux distribution to the randomly generated v_{rand}. More specifically, $\delta_{irr(i)} = \delta^+_{irr(i)} - \delta^-_{irr(i)} = v_{rand(i)} - v_{irr(i)}$, $i \in I_R$, acts over the set of irreversible reactions (I_R) and $\delta_{for(i)} = \delta^+_{for(i)} - \delta^-_{for(i)} = v_{rand(i)} - v_{for(i)}$, $\delta_{back(i)} = v_{rand(i)} - v_{back(i)}$, $i \in R_R$, over the set of reversible reactions (R_R). Note that both δ_{irr}, δ_{for}, are defined as the difference of two non-negative components, which enables us to formulate a linear objective function that renders (3.2) computationally tractable. In contrast, δ_{back} does not require this treatment since it always takes non-negative values (see Figure 3.2). This is because in the MILP (3.2), the stoichiometric matrix, S, corresponding to the GEM is first modified in the following way: we change the sign of the columns, as well as the entry in v_{rand}, corresponding to reversible reactions that were randomly assigned a negative flux value in v_{rand}

A

RegrEx$_{AOS}$

$\{v_{min}, v_{max}\} \leftarrow FVA$ (GEM)
niter = 1
flipped = {∅}
WHILE (niter ≤ nsamples)
  S' = S
  $v_{rand} \leftarrow rand(v_{min}, v_{max})$
  FOR each reaction i
    IF $v_{rand[i]} < 0$
      flip sign of i´th column of S'
      flip sign of i´th entry of v_{rand}
      flipped ← flipped∪ i
    END IF
    $v_{AO} \leftarrow OP_{32}$ (S', v_{min}, v_{max}, v_{rand}, v_{opt}, d)
    $v_{AO\,[flipped]} = - v_{AO\,[flipped]}$
  END FOR
  niter = niter +1
END WHILE

B

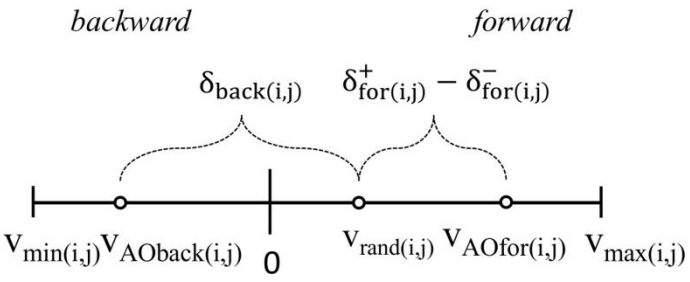


Figure 3.2. Pseudocode for RegrEx$_{AOS}$ and details of the treatment of reversible reactions. *(A) RegrEx$_{AOS}$ first finds the minimum, v_{min}, and maximum, v_{max}, allowable flux values through Flux Variability Analysis (FVA, see Methods, section 3.4) for each reaction in the GEM. It then repeats the following procedure until obtaining the required number of samples (nsamples). (i) Generate a random flux distribution, v_{rand}, in which each random flux value remains within the feasible range obtained before. (ii) Change the sign of the negative entries in v_{rand} and of the corresponding columns in the stoichiometric matrix. (iii) Generate an alternative optimal flux distribution, v_{AO}, that is closest to v_{rand} through the optimization program (OP) in (3.2), which takes the modified stoichiometric matrix, S', v_{min}, v_{max}, v_{rand}, the previous optimum RegrEx solution, v_{opt} and the data vector, d, as arguments. (iv) Change the sign of the entries in v_{AO} corresponding to the original negative entries in v_{rand}. (B) In RegrEx$_{AOS}$, reversible reactions are split into the forward and backward directions. The entries corresponding to reversible reactions in v_{rand} are always non-negative (since the sign is changed if negative), and fall in the range of the corresponding forward direction (since the sign of the associated column in S is changed accordingly). Hence RegrEx$_{AOS}$ can choose between $\delta^{+}_{for} - \delta^{-}_{for}$, quantifying the distance between v_{rand} and an optimal flux value in the forward direction, or δ_{back}, which measures the distance between v_{rand} and an optimal flux value in the backward direction. At the end of the optimization process (3.2), RegrEx$_{AOS}$ selects the direction of each reversible reaction that minimizes the overall distance to v_{rand}.*

In this manner, all reversible reactions in v_{rand} operate in forward direction (*i.e.,* are non-negative) which facilitates the optimization process. In addition, δ_{for} and δ_{back} are constrained to be mutually exclusive by the same binary variable, x, introduced to select only one of the directions in reversible reactions (*i.e.* either forward or backward). In this way, the optimization problem in (3.2) will select the direction of reversible reactions that minimizes the overall distance to v_{rand}. Finally, reversible reactions whose sign was originally changed in v_{rand} are altered back to their original directions and their sampled flux values are modified accordingly.

3.2.1.2 Alternative optimal solutions in network-centered approaches: the case of CorEx

In this section, we analyze the alternative optimal solutions of CorEx, a method that we designed in this study to represent the network-centered approaches. In a general sense, network-centered approaches first partition the set $R = C \cup P$ of reactions in the original GEM into a core set, C, that must be present in the final context-specific model, and a non-core set, P, which does not necessarily have to be in the final model. These approaches find then a subset $P_A \subseteq P$ of non-core reactions that renders C consistent, *i.e.*, all reactions in the core are able to carry a non-zero flux in at least one steady-state solution. The final context-specific subnetwork is then defined as $R_A = C \cup P_A$. Some approaches, like MBA (Jerby et al., 2010), mCADRE (Yuliang Wang et al., 2012) and FastCORE (Vlassis, Pacheco, et al., 2014), aim at minimizing the size of P_A, as to obtain a parsimonious final model. In contrast, CORDA (Schultz & Qutub, 2016) relaxes the parsimony condition as a way to prevent eliminating important reactions for a given context. In this respect, CorEx aims at obtaining a parsimonious model, although, as shown in the following, it can be easily adapted to allow increasing the size of P_A if desired.

CorEx follows the MILP displayed in (3.3), which minimizes the number of reactions with non-zero flux in P while constraining all reactions in the core to carry at least a small positive flux (ϵ in constraints 2-3). This is achieved by minimizing the norm (Z in (3.3)) of the vector, x, of binary variables (constraints 4-7) which selects the set P_A that renders the MILP feasible. Note that the selected non-core reactions are forced to carry a small positive flux (constraints 5, 7) to guarantee that they are active in the final context-specific model. Finally, like in RegrEx, reversible reactions are split into the forward and backward directions, to operate only with non-negative flux values. In addition, another vector of binary variables, y in constraints 8-9 of (3.3), is introduced to select the direction of reversible reactions (*i.e.*, imposing $v_{for} > XOR\ v_{back} > 0$, when the reaction is selected to be active).

$$
\begin{aligned}
&\min_{\substack{v=[v_{irr};v_{for};v_{back}]\in\mathbb{R}_0^{r+},\\ x=[x_{irr};x_{rev}]\in\{0,1\}^{|P|}\\ y\in\{0,1\}^{rev}}} \quad \| x \|_1 \\
&s.t. \\
&1.\ S_{ext} v = 0 \\
&\left.\begin{aligned} 2.\ & v_{irr(i)} \geq \varepsilon \\ 3.\ & v_{for(i)} + v_{back(i)} \geq \varepsilon \end{aligned}\right\}, \quad i \in C \\
&\left.\begin{aligned} 4.\ & v_{irr(i)} - x_{irr(i)} v_{\max} \leq 0 \\ 5.\ & v_{irr(i)} - x_{irr(i)} \varepsilon \geq 0 \\ 6.\ & (v_{for(i)} + v_{back(i)}) - x_{rev(i)} v_{\max} \leq 0 \\ 7.\ & (v_{for(i)} + v_{back(i)}) - x_{rev(i)} \varepsilon \geq 0 \end{aligned}\right\}, \quad i \in P. \\
&8.\ v_{for} + y v_{\max} \leq v_{\max} \\
&9.\ v_{back} - y v_{\max} \leq 0
\end{aligned}
\tag{3.2}
$$

To identify alternative optimal CorEx extracted networks, we developed the MILP displayed in (3.4). The general idea behind the optimization problem in (3.4) is to find the most dissimilar context-specific network, $R_{A*} = C \cup P_{A*}$, to a previously found optimal R_A, that maintains the set C consistent. Namely, it maximizes the number of differences between the reactions contained in P_A and P_{A*}. Note that the optimization problem in (3.4) inherits constraints 1-9 from (3.3), and incorporates three new constraints. Constraint 10 guarantees that the cardinality of P_{A*} equals that of the previous optimal P_A in (3.3). Constraint 11 introduces two additional binary variables, δ^+, δ^-, which measure the mismatches between the vectors x, selecting the reactions in P_{A*}, and the optimal vector x_{opt}, selecting the reactions in P_A and previously found by the optimization problem in (3.3). Finally, constraint 12 is added to impose a δ^+ *XOR* δ^- relationship to avoid the trivial optimal solution in which $\delta^+ = \delta^-$,

$$
\begin{aligned}
&\max_{\substack{v=[v_{irr};v_{for};v_{back}]\in\mathbb{R}_0^{r+},\\ x=[x_{irr};x_{rev}],\delta^+,\delta^-\in\{0,1\}^P\\ y\in\{0,1\}^{rev}}} \quad \| \delta^+ + \delta^- \|_1 \\
&s.t. \\
&1-9.\ (3.3) \\
&10.\ \| x \|_1 = Z \\
&11.\ x + \delta^+ - \delta^- = x_{opt} \\
&12.\ \delta^+ + \delta^- \leq 1.
\end{aligned}
\tag{3.3}
$$

However, besides CorEx, the optimization problem (3.4) can be used to generate alternative optimal networks to other network-centered approaches. We just need to set x_{opt}, in constraint 11, to be the optimal x vector of the particular approach under study; in addition, we need to update Z, in constraint 10, to the corresponding number

of non-core reactions added by this approach (*i.e.*, the size of P_A). Note that x_{opt} can be easily constructed from the set P_A, which is derived from a particular context-specific model. In addition, a similar constraint to the constraint 10 of (3.4), namely $\|x\|_1 \geq Z_{lb}$, may be included in (3.3), as a lower bound to its objective function, where $Z^* \leq Z_{lb} \leq R$, and Z^* is the unconstrained optimum of (3.3). It is in this manner that CorEx allows relaxing the parsimony condition, as commented before, although in this study we did not constrain the CorEx optimum.

Noteworthy, the main advantage of using the optimization problem (3.4) to obtain alternative optimal networks lies in its MILP formulation. This is because, with the exception of CorEx, which also relies on a single MILP, all existing network-centered approaches require iteratively solving a convex optimization problem. For instance, the linear programs behind the searching for sparse modesin FastCORE (Vlassis, Pacheco, et al., 2014), or the ones behind the flux balance analysis, iterated over each reaction of the GEM, in CORDA (Schultz & Qutub, 2016). Alternative optima may arise in each one of these iterations, thus exploring the alternative optima space in each case would require an extensive computational effort. In contrast, we circumvent this problem with the optimization problem in (3.4) by analyzing the alternative solutions of a single MILP. However, this optimization only generates a single, maximally different, alternative optimal network.

To generate a sample of alternative networks, here we applied the optimization problem in (3.4) in an iterative way. We first used (3.4) to obtain a maximally different network to a given optimal context-specific network, and then repeated this process of feeding (3.4) with the successively generated alternative networks until no additional one was found. At that point, we randomly perturbed the last network by changing the state (active or inactive) of 1% of the reactions, and repeated this process until no additional network was found (an implementation of the procedure is provided in File S3.1). We note that with this iterative process, which we term the AltNet procedure, we do not guarantee an exhaustive enumeration of all maximally different alternative networks. However, as shown in the next section, it sufficed to illustrate the variety found across optimal context-specific extracted networks in this study.

Finally, we use the AltNet procedure to analyze the alternative optima space of CorEx, FastCORE and CORDA. In the latter case, however, the optimization problem in (3.4) had to be slightly modified. The reason for the modification is that CORDA divides the reactions in the GEM into four categories, in contrast to CorEx and FastCORE, where only the core, C, and the non-core set, P, are considered. Concretely, reactions are separated into three groups based on experimental evidence: reactions with *high* (HC), *medium*, (MC) and *negative* (NC) confidence, and an additional group collecting the remaining reactions (OT) in the GEM, for which experimental evidence is not available. In this case, the group HC corresponds to the core set of reactions (*i.e.,* all reactions in HC must be included in the final model), and the other three groups constitute the non-core set P, although reactions in MC are

preferentially added over NC and OT reactions. To account for the different reaction groups, we partitioned the vector *x* in (3.4) into the sets of MC, NC and OT reactions, and evaluated constraint 10 for each of the three sets. In this manner, we guaranteed that an alternative optimal network contained, besides all HC reactions, the same number of MC, NC and OT reactions than the original CORDA optimum.

3.2.2 Evaluation of alternative optima: Case studies

Here, we illustrate the ambiguity found during the extraction of context-specific flux distributions and metabolic networks due to the alternative optima. To this end, we apply the methods described in the previous section to two case studies: a leaf-specific scenario, the model plant *Arabidopsis thaliana,* and a human, liver-specific reconstruction. In the first case, we used the AraCORE model, which includes the primary metabolism of *Arabidopsis thaliana* (Arnold & Nikoloski, 2014), and a leaf-specific gene expression data set, obtained from (Booker, Burkey, Morgan, Fiscus, & Jones, 2012) (Methods, section 3.4). In the second case, we employed Recon1, a well-established human metabolic model (Duarte et al., 2007). Moreover, we considered two different core sets of reactions that were identified as liver-specific by experimental evidence (taken from [19] and [20]), and upon which the liver reconstructions were built. In addition, we reduced the original metabolic models by taking only the consistent part of them. The resulting models are termed here Recon1red and AraCOREred, and contain a total number of 2469 and 455 reactions, respectively (see Methods, section 3.4, for details).

We first analyzed the alternative optima space of $\text{RegrEx}_{\text{LAD}}$—as a representative of a flux-centered approach—and evaluated the ability of the ℓ_1-regularization of $\text{RegrEx}_{\text{LAD}}$ to reduce this space. To this end, we focused on the leaf-specific scenario; however, we also applied these methods to the liver-specific scenario, to verify if our main conclusions held in the case of a larger genome-scale model. We then applied CorEx, a network-centered representative, to extract and analyze the alternative optima for the leaf- and the liver-specific reconstructions, and compare its performance with that of FastCORE [19], a well-established approach. In addition, we evaluated the alternative optimal liver-specific networks generated by CORDA, a recently published approach [20]. Finally, we also investigated the alternative optima of iMAT to the leaf- and liver-specific scenario with both, the original approach proposed in [16] and our own complementary method.

3.2.2.1 Alternative $\text{RegrEx}_{\text{LAD}}$ optima during leaf-specific data integration

After applying $\text{RegrEx}_{\text{LAD}}$ with $\lambda = 0$, we obtained an optimal, leaf-specific flux distribution. We then applied $\text{RegrEx}_{\text{AOS}}$ to evaluate the alternative optima space of the previously obtained optimum. The results from this evaluation confirmed the existence of an alternative optima space for $\text{RegrEx}_{\text{LAD}}$. However, the variability of the fluxes at the optimal objective value was not uniform across different reactions.

As expected, data-orphan reactions exhibited more broadly distributed flux values at the alternative optima than data-bounded reactions. We quantified this property by the Shannon entropy (Methods, section 3.4), as a measure of uncertainty of flux value prediction associated to a data integration problem. In this sense, data-orphan reactions showed a larger mean entropy value of 1.64 in comparison to the value of 0.95 found for the data-bounded reactions (one-sided ranksum test, p-value = 1.95×10^{-5}). However, we found reactions with particularly low or high entropy values in both sets, data-bounded and data-orphan (Table S1).

This last observation suggests that reactions with low entropy values may be of special importance under the leaf-specific metabolic state. On the other side, high entropy values suggest that the corresponding reactions could operate more freely in the leaf context. For instance, we found that the majority of transport reactions showed large entropy values, in accord with the fact that most transport reactions are data-orphan. Nevertheless, there were some transport reactions with particularly low entropy values, such as: the *TP/Pi translocator* (reaction index 327 in AraCOREred, $H = 0.07$) interchanging glyceraldehyde 3-phosphate and orthophosphate between the chloroplast and cytoplasm, the *P5C exporter* (index 363, $H = 0.01$) exporting 1-Pyrroine-5-carboxylate from mitochondria to cytoplasm and the *ADP/ATP carrier* (index 320, H = 0.01), interchanging ATP and ADP also between mitochondria and cytoplasm. For a comparison, the highest entropy value in the rank is $H = 2.92$, corresponding to the *Proline uniporter* (see the complete list in Table S3.1). Therefore, the leaf data integration constrains these transport reactions to take a small range of different flux values due to the network context in which they operate, since they are not directly bounded by experimental data. This observation is contrasted by the high entropy values that these same three reactions when no experimental data are integrated, *i.e.*, when a similar sampling procedure is performed in which only mass balance and thermodynamic constraints are imposed (Methods, section 3.4). In this case, all three entropy values are markedly larger ($H > 2$, Table S3.1).

We next focused on the entropy values of reversible reactions in the AraCOREred model. Reversible reactions in a GEM usually correspond to reactions for which no thermodynamic information is available (leaving aside the set which is known to operate close to equilibrium). Therefore, it would be informative to evaluate whether integrating context-specific experimental data in a GEM could be used to fix the direction of such reactions. Interestingly, we found that a large proportion (75.81%) of the reversible reactions carrying a non-zero flux (including data-orphan) had a fixed direction, either forward or backward, in the alternative optima (Table 3.1). This finding indicates that, even though there is variation in the flux value of reversible reactions, integration of expression data can determine their direction in a given context. Therefore, the proposed approach and findings provide valuable information on how metabolism could be operating under the particular condition.

Leaf	***λ = 0***	***λ = 0.1***	***λ = 0.3***	***λ = 0.5***
H_{Data}	73.17	71.34	81.77	65.46
H_{Orphan}	86.82	62.18	59.97	36.50
H_{Total}	159.99	133.52	141.74	101.95
$\overline{H}_{Total}$	1.23	1.03	1.09	0.78
Fixed$_{Rev}$ (%)	75.81	75.81	80.95	98.18
Liver	**λ = 0**	**λ = 0.1**	**λ = 0.3**	**λ = 0.5**
H_{Data}	817.22	789.37	763.68	780.87
H_{Orphan}	810.79	658.66	488.31	310.21
H_{Total}	1628.14	1448.04	1251.99	1091.08
$\overline{H}_{Total}$	1.20	1.07	0.92	0.80
Fixed$_{Rev}$ (%)	61.78	60.31	62.41	52.09

Table 3.1 Summary of the alternative optima space of RegrEx$_{LAD}$ for two case studies, Leaf and Liver, and four values for the parameter λ. *For the analyzed sequence of increasing λ-values, the table includes: The sum of entropy values for the subset of data-bounded, H_{Data}, and data-orphan, H_{Orphan}, reactions, as well as for all reactions, H_{Total}, the mean entropy value across all reactions, $\overline{H}_{Total}$, and the proportion of reversible reactions with fixed direction in the alternative optima sample, $Fixed_{Rev}$.*

3.2.2.2 Effect of regularization on the alternative optima space

We next evaluated the RegrEx$_{LAD}$ alternative optima space for a sequence of increasing λ-values. This was motivated to test whether the inclusion of ℓ_1-regularization, besides imposing sparsity in optimal flux distributions, could also reduce the variability found in individual reaction flux values across the alternative optima space. This property could serve as a way to decrease the uncertainty, as measured by the Shannon entropy, associated to a context-specific data integration problem. To this end, we first applied RegrEx$_{LAD}$ on AraCORErd and the same leaf data set, but using three increasing λ-values ($\lambda_1 = 0.1$, $\lambda_2 = 0.3$ and $\lambda_3 = 0.5$). We then applied RegrEx$_{AOS}$ to sample the alternative optima space of each of the three RegrEx$_{LAD}$ data integrations.

We found that the entropy tended to decrease with increasing λ-values, although the effect was more pronounced for the data-orphan reactions (Table 3.1, Figure 3.3).

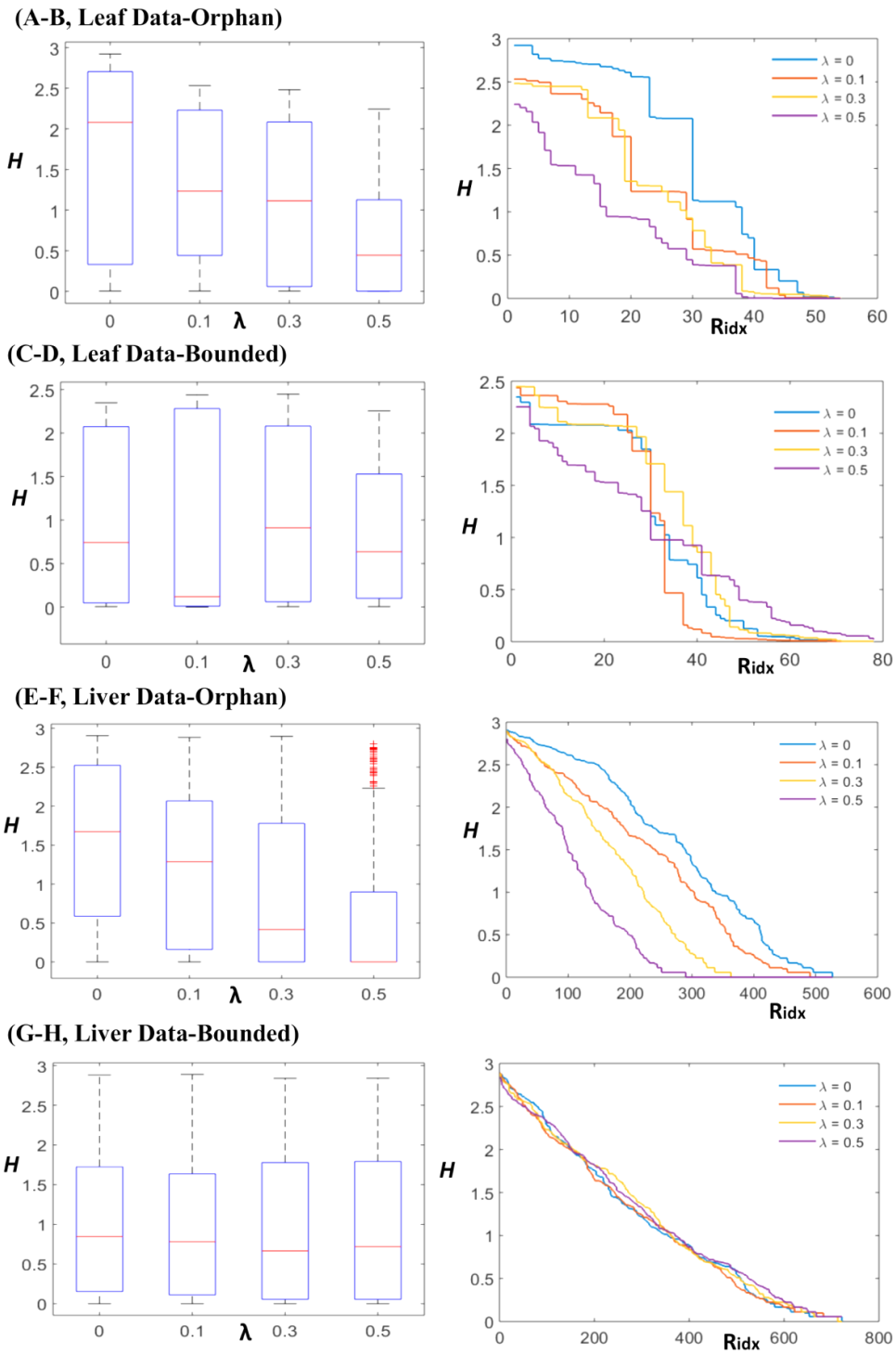


Figure 3.3. Effect of regularization on the alternative optima space of RegrEx$_{LAD}$. *The effects of regularization are presented, for the two case studies, by depicting the box plots of the distributions of Shannon entropy values, H. The distributions are partitioned into the set of data-orphan (A and E, for leaf and liver, respectively) and data-bounded reactions (C and G, for leaf and liver, respectively) across increasing λ-values. Median values, represented by red lines, decrease monotonically only in data-orphan reactions (bottom and upper edges in the box plots indicate the 25th and 75th percentile, respectively). Additionally, the individual entropies for each data-orphan (B and F, for leaf and liver, respectively) and data-bounded (D and H, for leaf and liver, respectively) reaction are also presented in decreasing order for the four λ–values (reactions with H < 10^{-3} are omitted). In data-orphan reactions, all distributions with λ > 0 fall below the corresponding to λ = 0 (without regularization, depicted in blue), which is not the case in data-bounded reactions.*

For instance, the sum of entropy values among data-orphan reactions decreased from a value of H_{Orphan} = 86.82 for λ = 0, to H_{Orphan} = 36.50 with λ = 0.5. In contrast, for the data-bounded reactions, it only decreased from a value of 73.17 with λ = 0 to 65.46 with λ = 0.5, and even led to a transient increase at λ = 0.3 (Table 3.1, Figure 3.3). These findings suggest that the inclusion of regularization can reduce the uncertainty associated to a context-specific data integration problem. Naturally, there is a trade-off between decreasing uncertainty and increasing sparsity of the obtained models, since greater λ-values also produce smaller models that may exclude reactions that are relevant to a particular context (Figure 3.4). However, a mild regularization (λ = 0.1) already had a substantial effect in reducing the uncertainty of the $\text{RegrEx}_{\text{LAD}}$ data integration in this analysis. Specifically, it decreased the total model entropy, defined as the sum of entropy values over all reactions, by 16.54% (from a value of H_{Total} = 159.99 for λ = 0, to H_{Total} = 133.52 with λ = 0.1, Table 3.1).

Finally, we focused on the effect that regularization had on reversible reactions. We found that the number of reversible reactions with fixed direction increased monotonically with increasing λ-values (Table 3.1). Hence, this finding suggests that a mild regularization can further constrain the direction in which a reversible reaction can proceed under a particular metabolic context.

3.2.2.3 The $\text{RegrEx}_{\text{LAD}}$ alternative optima in the liver-specific case

We next analyzed the alternative optima space of $\text{RegrEx}_{\text{LAD}}$ in the liver scenario. Specifically, we focused on evaluating whether the qualitative results obtained in the leaf context remained unchanged when using Recon1red, a larger genome-scale model. To this end, we used a liver-specific and publicly available gene expression data set (Krupp et al., 2012), and mapped it to the reactions in Recon1red following the same procedure as in the leaf-scenario (Methods, section 3.4). Obtaining samples in a larger model is more challenging, due to the increased computational time required to solve the MILP of (3.2). Therefore, we restricted our sample to 100 random points for each of the four λ-values evaluated here, as to avoid an excessively large computational time (the total sample time remained under 41 hours, see Methods, section 3.4, for details).

We observed a general qualitative agreement between the leaf and the liver scenarios throughout the increasing λ sequence (Figure 3.3, E-H). More specifically, data-orphan reactions showed a monotonic decrease in their median entropy values; however, this effect was less apparent in the case of data-bounded reactions. Specifically, although the total entropy values of data-bounded reactions tended to decrease with increasing λ, with the exception of λ = 0.5 (Table 3.1), these differences were not significant (one-sided ranksum test, α = 0.05). However, we observed marked differences when looking at the proportion of fixed reversible reactions. In general, this fraction was smaller in the liver scenario, 61.78% *versus* 75.81% with λ = 0 (Table 3.1), and, in contrast to the leaf case, it did not show an increasing trend with increasing λ-values. We conclude that, while the sample size was smaller than that in the leaf case, these results again suggest that a mild ℓ_1-regularization of $\text{RegrEx}_{\text{LAD}}$ can be of help in reducing the ambiguity of context-specific flux values.

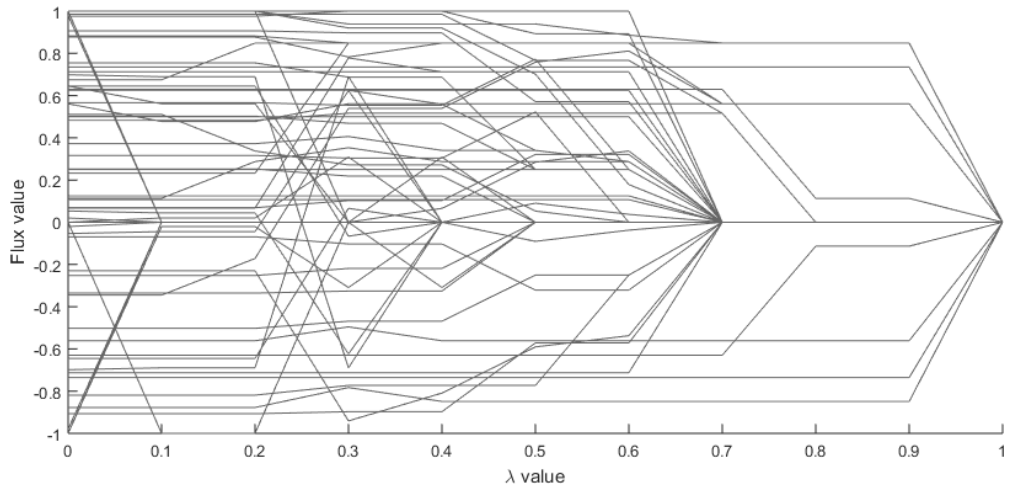


Figure 3.4. The RegrEx$_{LAD}$ solution path through a sequence of increasing λ-values. *A sequence of optimal solutions (i.e., flux distributions) to the leaf-specific RegrEx$_{LAD}$ integration problem is presented. The sequence begins with λ = 0 (i.e., no regularization) and ends with λ = 1, which is the value for which all fluxes are shrunk to 0. Flux distributions get sparser with increasing λ values. In addition, the total entropy of the alternative optima tends to decrease with increasing λ (Figure 3.3). This indicates the existence of a trade-off between sparsity and entropy reduction. In this study, a mild regularization (λ = 0.1) seems sufficient to substantially reduce the total entropy value while preventing flux distributions to become too sparse (i.e., in which important reactions for a given context may be excluded).*

3.2.2.4 Alternative optima in leaf- and liver-specific metabolic networks

We first applied CorEx and FastCORE to reconstruct two leaf-specific networks, Leaf$_{CorEx}$ and Leaf$_{FastCORE}$. To this end, we used the AraCORErred model and a core set of 91 reactions, which was previously obtained by considering reactions for which the associated gene expression data had a value greater than the 70th percentile (Methods, section 3.4). Both Leaf$_{CorEx}$ and Leaf$_{FastCORE,}$ contained the core set and were consistent, *i.e.,* all reactions were unblocked. However, we noticed that Leaf$_{CorEx}$ was more compact than Leaf$_{FastCORE}$, containing 236 versus 254 non-core reactions, respectively (Table 3.2). We next reconstructed the two liver-specific networks in a similar way. To this end, we used the Recon1red model, and the core set of 1069 reactions defined in the original FastCORE publication (Vlassis, Pacheco, et al., 2014). In this case, CorEx added 593 non-core reactions to the core set, obtaining the liver-specific reconstruction Liver$_{CorEx}$. FastCORE, on the other hand, it added 677 non-core reactions to generate Liver$_{FastCORE}$. Hence, CorEx was able to extract a more compact liver-specific network, resembling the behavior found in the leaf-specific case. After obtaining these context-specific metabolic reconstructions, we searched for alternative optimal networks to all of them, using the AlterNet procedure described in the previous section. To quantify the uncertainty of the leaf- and liver-specific reconstructions, we looked at the number of reaction mismatches between all

pairs of alternative networks in each case (computed as the Hamming distance, see Methods, section 3.4). This metric was normalized by the total number of reactions in each metabolic model to allow fair comparison between the two case studies.

	P	***#models***	$M_R max$	$\overline{M_R}$ ***(CV)***	***p-value***
$Leaf_{CorEx}$	236	61	52 [22%]	29.03(0.29)	0
$Leaf_{FastCORE}$	254	201	118 [46.5%]	66.76(0.54)	
$Liver_{CorEx}$	593	4	156 [26.3 %]	108.33(0.37)	0.0022
$Liver_{FastCORE}$	677	100	398 [58.8%]	247.93(0.46)	
$Liver_{CORDA}$	1527	104	992	545.22(0.42)	0
$Liver_{CORDAtest}$	1527	18	860	389.40(0.48)	

Table 3.2 Summary of the alternative optima space of the evaluated network-centered methods. *This table summarizes the results of the evaluation of the CorEx alternative optima space. It includes the number of added non-core reactions,* P*, the maximum,* $M_R max$ *(within brackets the percentage of reaction in* P*), and the mean number,* $\overline{M_R}$ *(CV stands for coefficient of variation), of reaction mismatches (*i.e., *Hamming distance) across the alternative networks for the leaf- and the liver-specific scenarios evaluated by two methods, CorEx and FastCORE. The last column displays the p-value resulted from a one-sided ranksum test comparing the distributions of Hamming distances between any pair of the alternative networks of CorEx and FastCORE (the null hypothesis states that the distribution generated by CorEx is bigger than that of FastCORE).*

We found marked differences between alternative optimal networks in both approaches and metabolic scenarios. In the case of $Leaf_{CorEx}$, alternative networks differed on average in 29 non-core reactions, with a maximum value of 52 reactions (22% of the added non-core reactions). In $Leaf_{FastCORE}$, networks differed on average in 66.78 reactions, and had a maximum number of 118 discrepant reactions (46.5%, Table 3.2). This situation was even worsened in the liver-specific reconstructions. Between alternative networks to $Liver_{CorEx}$, we found a maximum of 156 discrepant reactions among the 593 in the added non-core (26.3%), with an average of 108.3. In the case of $Liver_{FastCORE}$, the maximum number of discrepant reactions was as high as 398 out of the 677 (58.8%) added non-core reactions, with an average of 246.93 between alternative optimal networks (Table 3.2).

As a complementary analysis, we also determined the frequency of occurrence of every non-core reaction across the alternative optimal networks. In this manner, we could identify: (*i*) a set of non-core reactions that were always included, termed the active non-core set, (*ii*) a set of non-core reactions that were excluded from all alternative networks, termed the inactive non-core set, and (*iii*) a set of non-core reactions that were included in some of the networks, referred to as the variable non-

core set. In this case, we took the size of the variable non-core set as a measurement of the uncertainty of a context-specific network extraction; 28% and a 47% of the total non-core reactions were in the variable set in the cases of $Leaf_{CorEx}$ and $Leaf_{FastCORE}$. On the other hand, a 12% and a 58% were found in $Liver_{CorEx}$ and $Liver_{FastCORE}$, respectively (Figure 3.5, A-D).

The previous results quantify the structural differences among the generated alternative optimal networks. However, these structural differences do not consider which kind of reactions (*i.e.*, in which pathways in the GEM) are more or less frequent (*i.e*., ambiguous), in the alternative optima space. To address this issue, we assigned a score (between 0 and 1) to each metabolic pathway based on its representation in the active, variable or inactive non-core set. Specifically, the score represents the fraction of reactions of a given pathway that are assigned to a non-core subset with respect to the total size of the non-core set (Methods, section 3.4). Pathways with high score values in the active and inactive non-core are consistently over- and under-represented, respectively, among the alternative optimal networks. Therefore, these pathways should be more important (the opposite in the inactive non-core case) to maintain the core active and hence the assumed context-specific metabolic function. In contrasts, pathways with high-score values in the variable non-core tend to be represented only in certain alternative optimal networks, thus being more ambiguous in the context-specific reconstruction.

For instance, in the leaf scenario, we found among the pathways with highest score in the active non-core: the *Calvin-Benson cycle, light reactions* and *photorespiration.* All of these pathways showed a maximum score value of 1 in both cases $Leaf_{CorEx}$ and $Leaf_{FastCORE}$, which agrees with key roles of these pathways in a photosynthetic tissue. Additionally, alongside these photosynthetic pathways, we also found housekeeping pathways for the synthesis of AMP, CTP, GMP, UMP, Acetyl-coA or Fatty acid, among others, with the maximum score value in both cases. More interestingly, among the pathways with the highest scores in the variable set we also found primary pathways like the *Tricarboxylic acid cycle, Alanine synthesis,* the *Pentose Phosphate Pathway* and *Pyruvate metabolism*. However, we also found pathways that are usually linked to active photosynthetic tissues like *Starch* and *sucrose degradation* and *sucrose synthesis* (see Table S9 for a complete list containing the ranked pathways).

Moreover, in the liver scenario, we found typical liver-specific pathways like *Cholesterol Metabolism* and *Fatty acid oxidation* (Mitra & Metcalf, 2009) with the maximum score value in the active non-core in the case of $Liver_{CORDA}$. However, we also found a variety of other pathways with high scores in the variable non-core such as *CoA catabolism, ROS detoxification* or *Vitamin A metabolism*, which indicates that the variable non-core set contains a diverse set of metabolic functions that may be important to the canonical liver physiology (see Table S9 for a complete list of the ranked metabolic pathways).

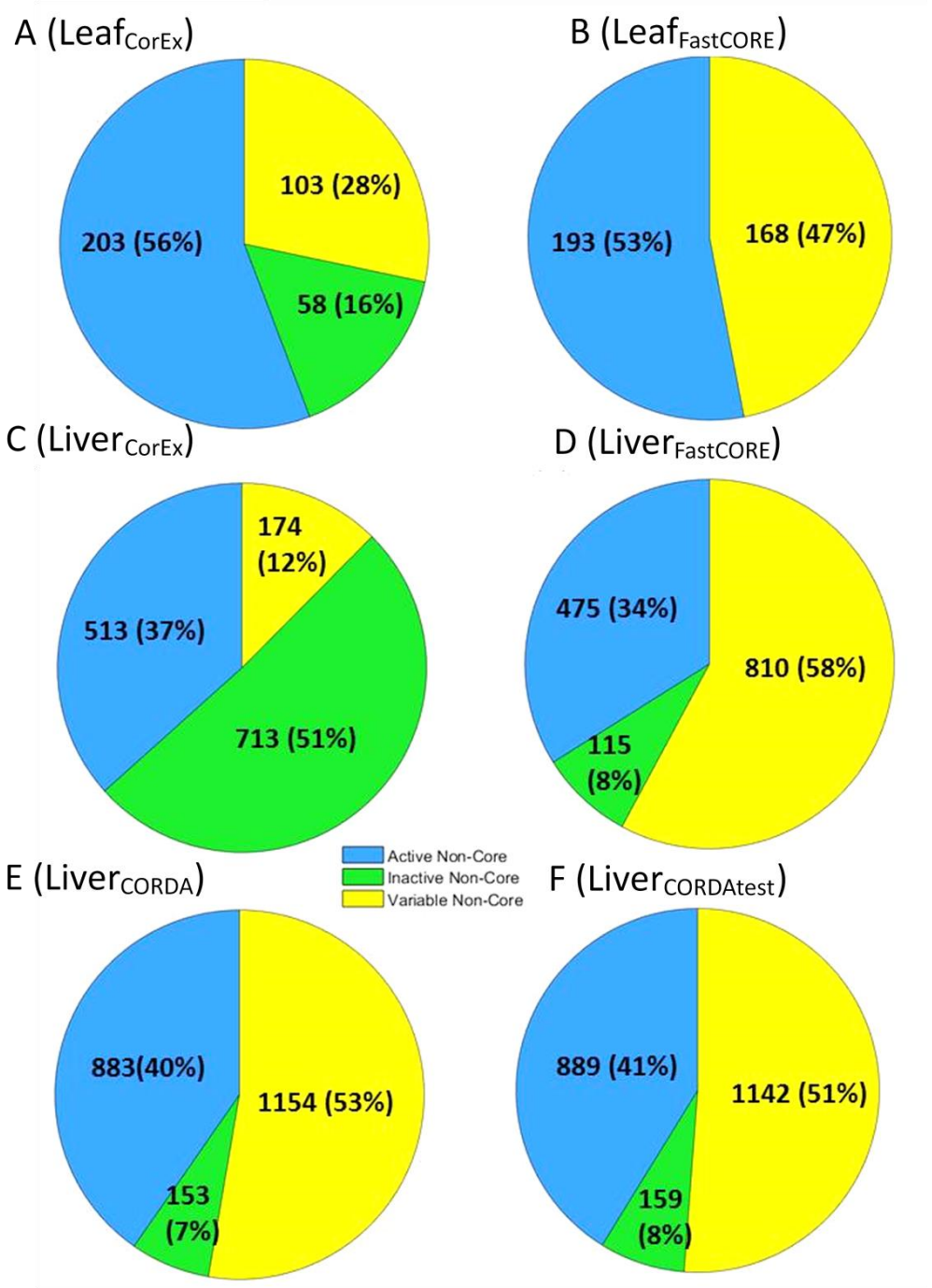


Figure 3.5. Alternative optima of CorEx and FastCORE context-specific network extractions. *The results are divided into the leaf-specific scenario for the CorEx (A) and FastCORE (B) alternative optima, and the liver-specific scenario, for CorEx (C), FastCORE (D) and CORDA without applying the metabolic test (E) and applying the metabolic test (F) to further constraint the alternative optima space (see main text). In all cases, non-core reactions are partitioned into the set that is always included in all alternative networks, (the fixed non-core set, in green), the set that is always excluded (excluded non-core, grey) and the variable non-core set (yellow) which is formed by reactions that are included in some of the alternative networks. In both, the leaf- and the liver-specific scenario, the alternative optima networks generated by CorEx contain a larger proportion of fixed non-core reactions and a smaller proportion of variable non-core reactions. These differences in behavior may be explained by the greater number of non-core reactions that are added by FastCORE, as compare to CorEx, in the optimal solution (see main text).*

Finally, we analyzed the alternative optima space of CORDA, a recently published network-centered approach (Schultz & Qutub, 2016). As explained in section 3.2.1, CORDA differs to CorEx and FastCORE in two ways. On one hand, CORDA does not aim at obtaining compact or parsimonious models, but rather emphasizes the metabolic functionality of the final context-specific reconstructions. On the other hand, CORDA considers four groups of reactions based on experimental evidence, out of which only one, the high confidence core set (HC), has to be fully included in the final model (thus being equivalent to the core set of CorEx and FastCORE). In this case, a suitable alternative optimal network must contain not only the entirety of the HC set, but exactly the same number of reactions added by CORDA in each one of the three remaining groups: the medium (MC) and the negative confidence (NC) groups, and the reactions without experimental data (OT). Therefore, it is reasonable to expect that this additional constraint may reduce the uncertainty of the CORDA reconstructions.

To test this idea, we searched for alternative networks to the CORDA liver reconstruction (here $Liver_{CORDA}$) provided in (Schultz & Qutub, 2016). $Liver_{CORDA}$ was obtained from Recon1 and experimental evidence from the Human Protein Atlas (Marx, 2014), and contains 279 HC, 369 MC, 11 NC and 1147 OT reactions. We used again our AltNet procedure, Recon1red (since blocked reactions, by definition, can never be included in a final network), and the classification of the reactions in the four groups also provided in (Schultz & Qutub, 2016). We were indeed able to find alternative networks to the original $Liver_{CORDA}$ with marked differences among them. Concretely, a maximum number of 992 discrepant reactions between two alternative networks, out of the total 1527 distributed among the MC, NC and OT groups (65%, Table 3.2), with a mean number of 545.22. Similarly, 51% of the non-core reactions (MC, NC and OT) in Recon1red were assigned to the variable non-core set (Figure 3.5, E).

The examples presented here show that the context-specific reconstructions are more ambiguous than specific, especially in the human liver scenario. This latter case is of special concern, given the implications of obtaining accurate context-specific reconstructions in biomedical research. In fact, most, if not all, of the network-centered approaches have focused on human metabolism (Jerby et al., 2010; Schultz & Qutub, 2016; Vlassis, Pacheco, et al., 2014; Yuliang Wang et al., 2012). There are ways, however, to cope with this ambiguity or uncertainty of context-specific reconstructions. For instance, as commented before, CORDA aims at obtaining functional reconstructions. In fact, the authors in (Schultz & Qutub, 2016) tested the capability of the $Liver_{CORDA}$ model to conduct a basic set of liver metabolic functions, including aminoacid, sugar and nucleotide metabolism.

We evaluated the alternative $Liver_{CORDA}$ models with the same metabolic test (Methods, section 3.4), and extracted the subset that passed it. Among these networks, we found that the number of discrepancies and the size of the variable non-core were significantly reduced, as compared to the total set of alternative networks (Table 3.2,

Figure 3.5, E-F). This is not surprising, since requiring the alternative networks to fulfill certain metabolic functions indirectly imposes an additional constraint to the optimal solution. On the other hand, this additional constraint can also be realized by augmenting the core set, as to guarantee that certain key reactions are present in the final context-specific network. This relates to an additional way to reduce the ambiguity of the reconstruction. In the case studies evaluated here, we found that the CorEx alternative networks tended to be more similar among each other than that of FastCORE or CORDA, as quantified by the (normalized by non-core size) number of discrepancies (Table 3.2). These differences may be explained by the number of non-core reactions included in the optimum: CorEx obtained more compact models than FastCORE in the Leaf- and the Liver-specific case. This imposes a more stringent constraint when searching for alternative optimal networks. However, there is a tradeoff between model parsimony and functionality. In fact, the $Liver_{CorEx}$ model was not able to pass the metabolic function test, while $Liver_{FastCORE}$ was able to pass it. In this particular case, $Liver_{CorEx}$ did not contain the 9 basal exchange reactions (Methods, section 3.4) required to perform the metabolic functions in the test. However, including these 9 reactions in the liver core set sufficed to generate a $Liver_{CorEx}$ model that passed the test.

The analysis of the alternative optima space can be employed to cope with the ambiguity of a context-specific network reconstruction. Notably, the authors of EXAMO (EXploration of Alternative Metabolic Optima) (Rossell et al., 2013) proposed a first step in this direction. In this case, EXAMO first generates a sample of alternative optimal flux distributions of iMAT (Shlomi et al., 2008). It then focuses on the activity state of each reaction across the sample, for which it binarizes the flux values through the usage of an arbitrary threshold value. A reaction is included in the *High Frequency Reaction* (HFR) set if it is active throughout the alternative optima sample. Finally, EXAMO uses the HFR set as a core set to MBA (Jerby et al., 2010), a network-centered method, which reconstructs the minimal network that renders the HFR set consistent. EXAMO directly addresses the problem of alternative optima. However, the final context-specific model is again subject to the effects of alternative optima, since a set of alternative networks, all containing the HFR set as a core, could be found for the MBA method.

A possible way to circumvent this problem in the case of iMAT could be the following: *i*) similar to EXAMO, obtain samples of alternative optimal flux distributions, binarize flux values and rank the reactions according to the number of times that they appear as active in the sample, *ii*) include the reactions that are always active (the HFR set) in a core set and the rest in a non-core set, and *iii*), add non-core reactions in decreasing order of frequency until consistency of the core is reached. In this manner, this ranking provides a way to select which non-core reactions are included in the final model. This idea parallels that of mCADRE (Yuliang Wang et al., 2012), although in the latter, reactions are ranked following an heuristic approach that considers experimental evidence from several databases, which may be difficult

to obtain for certain metabolic contexts. Finally, to generate the sample of alternative optima flux distributions of iMAT, we propose a sampling method similar to $RegrEx_{AOS}$ that allows drawing arbitrarily large samples, as opposed to the one used in EXAMO which generates samples of restricted size. Details about this method, here called $iMAT_{AOS}$, can be found in Appendix S3.2.

In the case of the network-centered approaches here evaluated, establishing a ranking of non-core reactions could also be a way to deal with the ambiguity during network reconstructions. Non-core reactions that occur with high frequency in the alternative optima space should be preferentially included in the final network, while reactions with a low frequency should be discarded. To guarantee that the final network is consistent (*i.e.* the core set is active), non-core reactions could be again added in decreasing order of frequency to the core set until consistency is reached. Naturally, this requires the development of competent methods to sample the alternative space of network-centered approaches. In this sense, we consider our proposed AltNet procedure a first step towards this goal.

3.3 Conclusions

We analyzed the space of alternative optima resulting from the integration of context-specific data into GEMs. To this end, we evaluated a representative from the flux- and network-centered approaches. We selected RegrEx (Robaina Estévez & Nikoloski, 2015) as a representative of flux-centered approaches and CorEx, as a network-centered approach, proposed in this study. In addition, we adapted CorEx to obtain alternative optimal networks for FastCORE (Vlassis, Pacheco, et al., 2014) and CORDA (Schultz & Qutub, 2016), two state-of-the-art network-centered approaches. We compared the developed approaches and implemented tools on two illustrative case studies: (*i*) a medium size GEM of the primary metabolism of *Arabidopsis thaliana* (Arnold & Nikoloski, 2014) and a leaf-specific gene expression data set (Booker et al., 2012), and (*ii*) a larger GEM collecting a reconstruction of a human metabolic network (Duarte et al., 2007), two liver-specific core sets of reactions (Schultz & Qutub, 2016; Vlassis, Pacheco, et al., 2014) and a liver-specific gene expression data set (Krupp et al., 2012).

Our findings demonstrated the existence of a space of alternative optima for all evaluated approaches integrating context-specific data. Consequently, this space of alternative optima induces ambiguous context-specific reconstructions. In the case of flux-centered approaches, $RegrEx_{LAD}$ in this study, we proposed the usage of a mild regularization to mediate the uncertainty of the resulting context-specific fluxes. In network-centered approaches, our results showed the existence of markedly disparate alternative context-specific networks in CorEx, FastCORE and CORDA. A delicate balance between model parsimony and metabolic functionality seems key to reducing the ambiguity of the context-specific reconstructions. Additionally, an evaluation of the alternative optima space followed by a ranking of the reactions according to their

frequency may serve as a way to determine their context-specificity. On this line, we proposed the AltNet procedure to generate alternative optimal context-specific networks.

As a concluding remark, we acknowledge the utility of the existent experimental data integration methods, since they allow a fast and automated generation of context-specific flux distributions and metabolic networks. However, our findings indicated that the interpretation and further usage of their results warrant caution. Specially, since the existence of alternative optima is likely linked to the nature of the context-specific data integration problem, and thus is independent of the approach used. The latter claim is supported by our evaluation across qualitatively different approaches. We advocate the view that an analysis of alternative optimal solutions should be performed, whenever possible, if context-specific data are integrated in metabolic models. In the case of context-specific networks reconstructions, more reliable results could be obtained from subsequent careful knowledge-based curation.

3.4 Methods

This section contains the details about the implementation of the methods described in this study, the GEMs and context-specific data employed in the case examples, and the computation of the distance metric between alternative optimal networks. In addition to this section, the MATLAB code containing the entire workflow followed in this study can be found in the Supplementary Information.

3.4.1 RegrEx$_{LAD}$, RegrEx$_{AOS}$, CorEx and AltNet implementations

All optimization programs used in this study, (3.1-4) were implemented in MATLAB and solved using Gurobi (version 7.1) (Gurobi Optimization, 2017) on a desktop machine with an Intel Core i7-4790 @3.6 GHz processor and 16GB of RAM. We used default Gurobi parameter values except for: *i*) reduced feasibility tolerance to 10^{-9} when solving (3.3-4), *ii*) increased MIPGap parameter to 1% when solving the MILP in (3.2). All generated code with the implementations is available as Supplementary Information.

3.4.2 Metabolic model and gene expression data

A reduced version of the original AraCORE model (Arnold & Nikoloski, 2014) was used in this study: AraCORE contains 549 reactions and 407 metabolites assigned to four subcellular compartments, whereas the herein used version (AraCOREred) contains 455 reactions and 374 metabolites. The reactions that were removed correspond to exchange reactions that directly connect organelles to the environment (circumventing the cytoplasm), and were eliminated to avoid bias in the obtained flux distributions. AraCOREred can be found in the Supplementary Material.

Leaf-specific gene expression values were taken from (Booker et al., 2012), stored in the GEO database under the accession numbers GSM852923, GSM852924 and GSM852925 corresponding to *Arabidopsis thaliana* Col-0 lines with no treatment. The corresponding CEL files were normalized using the RMA (Robust Multi-Array Average) method implemented in the *affy* R package (Gautier, Cope, Bolstad, & Irizarry, 2004). In addition, probe names were mapped to gene names following the workflow described in (Moyano, Vidal, Contreras-López, & Gutiérrez, 2015), where probes mapping to more than one gene name are eliminated. Gene expression values were then scaled to the maximum value and mapped to reactions in the AraCOREred model following the included Gene-Protein-Reaction rules and a self-developed MATLAB function, *mapgene2rxn*, which is available in File S1. This process was repeated for the three samples in the dataset and mean values were taken as representative values to obtain the final leaf-specific data used in this study.

Liver-specific gene expression values were obtained from (Krupp et al., 2012), which is accessible under: http://medicalgenomics.org/ma_seq_atlas/download. In this case, we used the RPKM values corresponding to the liver (normal tissues). Since the RPKM values are already normalized we used them directly as input of the *mapgene2rxn* procedure, already described.

We removed blocked reactions from the original Recon1 model to get the Recon1red model used in this study. To this end, we performed a Flux Variability Analysis (see next section) and removed reactions with a maximum absolute flux, $|v_i| < 10^{-6}$. The Flux Variability Analysis was implemented in the MATLAB function *reduceGEM* which also extracted the reduced model, Recon1red, in a COBRA compatible MATLAB structure. The function is available in File S1.

3.4.3 Extreme flux values of the flux cone

The minimum and maximum allowed values of each reaction in AraCOREred were determined through Flux Variability Analysis (R. Mahadevan & Schilling, 2003). Although only the mass balance and the thermodynamic constraints were imposed (*i.e.*, no reaction was forced to take a fraction of a previously calculated optimal value). This was accomplished through the following linear program,

$$\begin{aligned} &\min/\max_{v} v_i, \quad \forall i \in v \\ &s.t. \\ &Sv = 0 \\ &v_{\min} \le v \le v_{\max} \ , \end{aligned} \tag{3.4}$$

which was implemented in MATLAB and solved with the Gurobi solver (version 6.04). The own-developed MATLAB function can be found in Supplementary Material under the name of *FVA*.

3.4.4 Sampling flux distributions from the flux cone

To evaluate to what extent the Leaf data integration affected the entropy values of the reactions in the AraCORErd model, we also sampled the space of feasible flux distributions (*i.e.*, the flux cone) when no experimental data was integrated. To this end, and to allow direct comparability of the results, the flux cone was sampled following a similar approach as in RegrEx$_{AOS}$: first, we generated a random vector of flux values, v_{rand}, within the minimum and maximum values obtained by regular Flux Variability Analysis. The closest flux vector v to v_{rand} within the flux cone was then obtained by minimizing the Euclidean distance between the two vectors. The following quadratic program was used to this end:

$$\begin{aligned} &\min_{v} \tfrac{1}{2}\left\| v - v_{rand} \right\|_2^2 \\ &s.t. \\ &Sv = 0 \\ &v_{\min} \le v \le v_{\max} \ . \end{aligned} \tag{3.5}$$

This procedure was iterated to obtain a sample of size n = 2000. After the sample was generated, we obtained the Shannon entropy values of the samples in the same way as when evaluating the alternative optima space of RegrEx$_{LAD}$ (described in the next section). The MATLAB function implementing this sampling procedure can be found in File S1 under the name *coneSampling.*

3.4.5 Quantification of the RegrEx$_{LAD}$ alternative optima space

The Shannon entropy of the sampled alternative optima distribution, H_i, was used to quantify the extent to which the flux values of a reaction, i, varied across the alternative optima space. It was calculated as follows:

$$H_i = -\sum_{k=1}^{n} f_{i,k} \log(f_{i,k}), \tag{3.6}$$

where $f_{i,k}$ represents the frequency (*i.e.,* number of counts relative to sample size) of the k interval in the distribution, for n = 20 equally spaced flux value intervals within the flux range of i. In addition, the total entropy of an alternative optima space, H_T, was defined as the sum of the entropies corresponding to the r reactions in AraCORErd, *i.e.*,

$$H_T = \sum_{i=1}^{r} H_{v(i)} \ , \tag{3.7}$$

and was taken as a measure of the total flux variability found in a particular alternative optima space.

3.4.6 Measuring the distance between alternative optimal networks

In the case of CorEx, we generated the set of alternative optimal metabolic networks from the set of sampled alternative optimal flux distributions. To this end, we first generated the binary vector representations of the flux distributions. The binary vector representations were generated by assigning a value of 1 to the entries corresponding to reactions with a flux value $v \geq 10^{-6}$, and 0 otherwise. This process was repeated for each sampled alternative optimal flux distribution. In addition, repeated vector representations were removed from the generated set. After the binary representations were obtained, we calculated the number of mismatches between any pair, *a,b*, of binary vectors, with $a \neq b$, *i.e.*, the Hamming distance

$$M_{R(a,b)} = \sum_{k=1}^{n} | a_{(i)} - b_{(i)} | . \tag{3.8}$$

In this way, we obtained a distribution of M_R values whose characteristics were reported and compared.

3.4.7 Generation of a ranked list of metabolic pathways

We computed a score, ranging between 0 and 1, to quantify the ambiguity found in individual metabolic pathways (subsystems in the GEM) across the space of alternative optimal networks. Concretely, the score of a pathway, *M*, represents the fraction of the reactions in the (total) non-core set, *P,* belonging to the pathway that are assigned to the active, variable or inactive non-core (thus producing a score value for each case). That is, in general,

$$S_x(M) = \frac{X_M}{P} \ , \tag{3.9}$$

where $X_M \in \{A_M, V_M, I_M\}$ represents the number of reactions assigned to *M* that are included in the active, variable or inactive non-core, respectively.

3.4.8 Functional testing of the liver-specific reconstructions

We performed the same metabolic test proposed in (Schultz & Qutub, 2016) and applied to the original Liver-specific CORDA reconstruction. This test consists of a list of metabolic tasks that a metabolic model has to perform, including parts of the aminoacid, sugar and nucleotide metabolism. Concretely, there a total of 48 metabolic tasks, divided into the production of different aminoacids from minimal metabolic sources and the excretion on urea (19 tasks), the ability to synthetize glucose from 21 different sources (including some aminoacids), and the production of all 5 nucleotides and nucleotide precursors (8 tasks). The details about these tasks can be found in the

original CORDA publication (Schultz & Qutub, 2016), while the MATLAB code of our implementation of this test is provided in File S3.1. In this study, we used the fraction of performed tasks as measure of the ability of a given liver-specific model to pass this test. For instance, the liver-specific model provided in (Schultz & Qutub, 2016) (under the name of liverCORDAnew), was able to pass 89.58% of the tasks (43 out of 48). In this study, however, we required to pass all tasks in the test to consider an alternative liver-specific network as functional. We realized that, in the liverCORDAnew model, some reactions were slightly different to the analogous reactions in the Recon1red model that we used throughout this study (likely due to different versions of the Recon1 model, which is periodically updated (King et al., 2016)). When we reconstructed our $\text{Liver}_{\text{CORDA}}$ model, using the same reaction identifiers in liverCORDAnew but extracting the reactions from our Recon1red version, we found that the generated model passed all metabolic 48 tasks in the test. Hence, for consistency of the results, we considered that all proper alternative optimal networks to $\text{Liver}_{\text{CORDA}}$ had to pass all 48 tasks as well.

Acknowledgments

SRE would like to thank the Max Planck Society and the International Max Planck Research School “Primary Metabolism and Plant Growth” for providing the funding.

Author contributions

Performed the experiments and analyzed the data: SRE. Developed the methods and wrote the code: SRE. Wrote the paper: SRE, ZN. Developed the original idea: SRE, ZN.

Chapter 4

Applying RegrEx to investigate guard cell metabolism

Published as:

Resolving the central metabolism of Arabidopsis guard cells

Semidán Robaina-Estévez[1], Danilo de Menezes Daloso[2,3], Youjun Zhang[2], Alisdair R. Fernie[2], Zoran Nikoloski[1]

[1]Systems Biology and Mathematical Modeling Group, [2]Central Metabolism Group, Max Planck Institute of Molecular Plant Physiology, Potsdam-Golm, Germany, [3]Departamento de Bioquímica e Biologia Molecular, Universidade Federal do Ceará, Fortaleza, Brasil.

Scientific Reports (2017), DOI: .10.1038/s41598-017-07132-9

Abstract

Photosynthesis and water use efficiency, key factors affecting plant growth, are directly controlled by microscopic and adjustable pores in the leaf—the stomata. The size of the pores is modulated by the guard cells, which rely on molecular mechanisms to sense and respond to environmental changes. It has been shown that the physiology of mesophyll and guard cells differs substantially. However, the implications of these differences to metabolism at a genome-scale level remain unclear. Here, we used constraint-based modeling to predict the differences in metabolic fluxes between the mesophyll and guard cells of *Arabidopsis thaliana* by exploring the space of fluxes that are most concordant to cell-type-specific transcript profiles. An independent ^{13}C-labeling experiment using isolated mesophyll and guard cells was conducted and provided support for our predictions about the role of the Calvin-Benson cycle in sucrose synthesis in guard cells. The combination of *in silico* with *in vivo* analyses indicated that guard cells have higher anaplerotic CO_2 fixation via phospho*enol*pyruvate carboxylase, which was demonstrated to be an important source of malate. Beyond highlighting the metabolic differences between mesophyll and guard cells, our findings can be used in future integrated modeling of multi-cellular plant systems and their engineering towards improved growth.

4.1 Introduction

The stomata, microscopic and adjustable pores on the leaf surface, directly control two of the most important parameters affecting plant growth: carbon dioxide (CO_2) uptake from the environment and transpiration (Hetherington & Woodward, 2003). Thus, knowledge of the processes involved in stomatal movement is fundamental to understanding plant growth, and may represent a route to optimizing crop yield under the increasingly challenging environmental conditions (Medeiros, Daloso, Fernie, Nikoloski, & Ara??jo, 2015; Misra, Acharya, Granot, Assmann, & Chen, 2015). Stomatal movement depends on variations of the volume of two highly specialized, kidney-shaped cells, the guard (G) cells, which juxtapose to form the pore. The variations in the volume of the G cells are the macroscopic result of an intricate network of molecular processes occurring at different hierarchical scales (Santelia & Lawson, 2016). G cells stand out from the rest of the epithelial tissue—in which they are embedded—not only for their particular shape, but also for the remarkable property of containing photosynthetically active chloroplasts. Rather than contributing to total leaf carbon fixation, it has been suggested that active chloroplasts may be linked to the particular energetic and metabolic requirements for adequate G cell functioning(Lawson, 2009). In contrast, carbon fixation is the primary task of the main photosynthetically active cells, the mesophyll (M) cells. Although G and M cells are physiologically differentiated, the close connection between stomatal aperture and photosynthetic efficiency likely involves a fine coordination between these two cell types (Lawson, Simkin, Kelly, & Granot, 2014; Santelia & Lawson, 2016).

G cell represents a multisensorial system that responds to endogenous and environmental signals. Therefore, understanding of the complex cellular processes behind stomatal movement requires a systems approach to integrate experimental data with mathematical description of the underlying mechanisms. In practice, however, a complete mathematical description of stomatal movement is challenging due to experimental challenges and the hierarchy at which the key processes take place. Nevertheless, several studies, focusing on the dynamical processes of stomatal aperture, have rendered promising results derived from small-scale kinetic models (Hills, Chen, Amtmann, Blatt, & Lew, 2012; Li, Assmann, & Albert, 2006; Minguet-Parramona et al., 2016). For instance, the OnGuard modeling framework has been instrumental for explaining the dynamics of stomatal movement (Z.-H. Chen et al., 2012; Hills et al., 2012; Lawson & Blatt, 2014; Medeiros et al., 2015). It is based on a system of ordinary differential equations modeling the relationships between the influxes and outfluxes of water and different inorganic and organic osmolytes, the membrane potential, and macroscopic variables such as: total guard cell volume, turgor pressure and stomatal aperture. OnGuard also provides a phenomenological description of the metabolic processes involving the main organic osmolytes: sucrose and malate. However, it remains silent with respect to a detailed description of the

genome-scale metabolic processes occurring in G cells, since they are out of the scope of the kinetic modeling approach.

Relatively little is known about the genome-scale metabolic differences between G and M cells, although they can provide key insights into the modulation of metabolism in the two cell types (Lawson & Blatt, 2014). A genome-scale description of the metabolic state of G cells would provide a valuable complement to the existing kinetic models (Li et al., 2006; Sun, Jin, Albert, & Assmann, 2014). To this end, one can use the advances in genome-scale, constraint-based modeling of plants (Arnold & Nikoloski, 2014; de Oliveira Dal'Molin, Quek, Palfreyman, Brumbley, & Nielsen, 2010; Mintz-Oron et al., 2012; Poolman et al., 2009; Seaver et al., 2014), which have facilitated testing of hypotheses concerning the (re)distribution of steady-state metabolic activity under various conditions (Nikoloski, Perez-Storey, & Sweetlove, 2015). Integration of cell-type-specific data in this modeling approach is important since direct measurements of metabolic activity at a systems level are currently infeasible(Blazier & Papin, 2012; D. R. Hyduke et al., 2013). Transcriptomics data have been successfully employed to derive activity patterns in context-specific metabolic networks across a variety of organisms, from prokaryotes to more complex eukaryotes (Colijn et al., 2009; Lewis, Cho, Knight, & Palsson, 2009; Shlomi et al., 2008), and are readily available for G and M cells (Aubry et al., 2016; Bates et al., 2012; Bauer et al., 2013; Leonhardt et al., 2004; R.-S. Wang et al., 2011; Yizhou Wang & Blatt, 2011; Yang, Costa, Leonhardt, Siegel, & Schroeder, 2008).

Constraint-based modeling with integration of transcriptomics data provides a way to conduct differential analysis of metabolic activity between G and M cells. However, using transcriptomic—or even proteomic—data as an indicator of metabolic state calls for further justification, since metabolism is downstream in the cellular hierarchical organization. Integration of transcriptomics data is generally justified by two arguments: (*i*) transcript levels are currently the only data type with genome-scale coverage among the alternatives, protein or metabolic flux measurements, and (*ii*) transcript are not meant as proportional proxies of metabolic activity, but are rather used to constrain the fluxes in the large -scale model. While such an approach provides the basis for genome-wide differential flux profiling, it faces the challenge of multiple alternative optima whereby metabolic predictions for the same context-specific data (Robaina Estévez & Nikoloski, 2017), i.e., different metabolic states equally fit the data. Therefore, a robust differential analysis between cell-specific metabolic states requires *a priori* evaluation of the alternative optima, as to avoid biased conclusions based on selecting a single optimal metabolic state as a representative.

The main contributions from our constraint-based modeling study based on integration of G- and M-specific transcriptomics data include the following: (*i*) anaplerotic carbon fixation by phosphoenolpyruvate carboxylase (PEPc) is an important contributor to the production of malate in G cells, (*ii*) transport of oxaloacetate (OAA) to the mitochondria followed by malate production and its export

to the cytosol is the main contributor to the cytosolic malate pool, (*iii*) G cells perform an active photophosphorylation comparable to that of M cells, but they differ in the production of NADPH, and (*iv*) sucrose synthesis is dominant in G cells due to the presence of a futile cycle, not due to starch degradation. Our results suggested that G cells have adapted their metabolism towards production of malate and NADPH. We then showed that the key modeling predictions were robust and were in line with data from an independent ^{13}C labeling experiment, performed under similar conditions to those of the transcriptomic data. Therefore, our study constitutes a first step towards a quantitative, genome-scale analysis of the metabolic adaptations of G cells, and paves the way to further extensions to obtain a complete understanding of G cell physiology, with possible applications to crop engineering.

4.2 Results and discussion

4.2.1 Computational workflow and rationale for model-driven predictions of differences in G and M cell metabolism

To arrive at cell-specific metabolic predictions, we integrated G- and M-specific gene expression data in a modified version of the AraCORE model (Arnold & Nikoloski, 2014), here named AraCORErred (Material & Methods, Supplementary Figure S1, Appendix S3.1, Supplementary File S1). Our approach is fundamentally data-driven, and we did not include any cell-specific metabolic constraints to avoid bias while minimizing the discordance between fluxes and associated transcript levels. In constraint-based metabolic modeling, a metabolic state is characterized by the steady-state flux values through the reactions in the system (Orth et al., 2010). In addition, a metabolite can be described by the sum of steady-state fluxes of reactions in which the metabolite participates. This sum of fluxes is referred to as a flux-sum (Chung et al., 2009), and quantifies the flux through the pool of the respective metabolite. Therefore, here we predicted reaction fluxes and metabolite flux-sums from flux distributions in concordance with the data, which we then employed to dissect the differences between G and M cell metabolism.

Since the optimum to this multidimensional optimization problem is often not unique, we considered the set of alternative optima to this minimization problem, *i.e.*, the set of flux distributions that are equally similar to the data. To this end, we first obtained a representative sample of the steady-state reaction fluxes and metabolite flux-sums from the optimal solution space. We then applied a Mann-Whitney test to the resulting distributions to assess if, in a given cell type, a particular reaction or a metabolite showed significantly greater flux or flux-sum value, respectively. In addition, we used a complementary approach in which the extreme alternative optimal flux values for each reaction in AraCORErred—*i.e.*, minimum and maximum flux values equally concordant with data—were computed and compared between G and M cells. We stress that a comparison of alternative optimal samples of flux values is

preferred over a comparison of extreme flux values at alternative optima. The reason is that the flux range alone is not sufficient to provide a robust comparison between the metabolic states of the two cell types. In fact, it may be the case for a reaction to have the same minimum and maximum alternative optimal flux values in both cell types, whereas the distribution of flux values can markedly differ. For instance, this is the case if in one cell type the distribution is skewed to the minimum flux value and, in the other cell type, to the maximum.

To facilitate the interpretation of the predictions, we also determined and reported the mean flux and flux-sum values of each distribution and their ratios between G and M cells (all flux and flux-sum values are expressed in arbitrary units). However, we would like to stress that the differential analysis of fluxes is based on comparison of distributions of data-compatible flux values, by employing the Mann-Whitney test, and not on the comparison of their means. Therefore, in this study we refer to a flux as differential if its respective distributions differ, although these distributions may have the same mean.

4.2.2 Interplay between the tricarboxylic acid cycle (TCA) cycle and PEPc in the synthesis of cytosolic malate

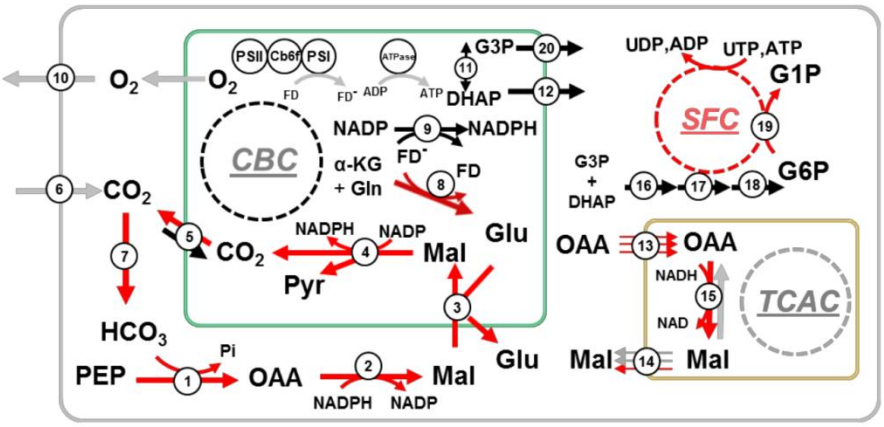
Malate has been repeatedly identified as a major osmoregulator controlling stomatal opening and closure (Fernie & Martinoia, 2009; Santelia & Lawson, 2016). Cytosolic carbon fixation by PEPc, followed by reduction of oxaloacetate (OAA) by cytosolic NADP-dependent malate dehydrogenase (NADP-MDH) may represent an additional source of malate in G cells, supported by ^{14}C and ^{13}C labeling experiments (Daloso, Antunes, et al., 2015; Gotow, Taylor, & Zeiger, 1988; Reckmann, Scheibe, & Raschke, 1990; Willmer & Dittrich, 1974) and by gene expression and enzyme activity measurements (Aubry et al., 2016; Bates et al., 2012; Bauer et al., 2013; W. H. Outlaw & Kennedy, 1978; Parvanthi & Raghavendra, 1997). These studies showed a high expression or activity of enzymes of anaplerotic fixation pathway in G cells (Daloso, dos Anjos, & Fernie, 2016). However, these findings have been countered by others claiming that malate content in G cells primarily depends on the supply from the surrounding M cells (Araújo et al., 2011; Nunes-Nesi et al., 2007; Penfield et al., 2012). Therefore, although there is a general consensus in considering malate a key regulator of stomatal regulation, its source in G cells remains unclear.

To delineate which pathway was the main contributor to shaping the pool of malate, we analyzed the relative contributions of cytosolic PEPc – NADP-MDH and the TCA cycle to malate production in G and M cells. This modeling strategy avoids setting a lower bound on non-zero malate uptake (from M cells) and allows an unbiased comparison of fluxes in the two cell types.

In comparison to M cells, the mean fluxes in G cells through the triad: Carbonic anhydrase (CA), PEPc and cytosolic NADP-MDH, were upregulated by a factor of ~

12, 12 and 2, respectively, that was significant when comparing the distributions of alternative optimal flux values (Figure 1A, Supplementary Table S1). These predictions demonstrated that the anaplerotic CO_2 fixation by PEPc plays a significant role in the production of malate in G cells. Since the average increase in the flux through NADP-MDH was smaller than the production of OAA by PEPc, it could be that a great part of the produced OAA is exported to the mitochondria and then converted into malate by the activity of NAD-MDH. To test this possibility, we next inspected the contribution of the TCA cycle to the production of cytosolic malate in G and M cells.

A



B

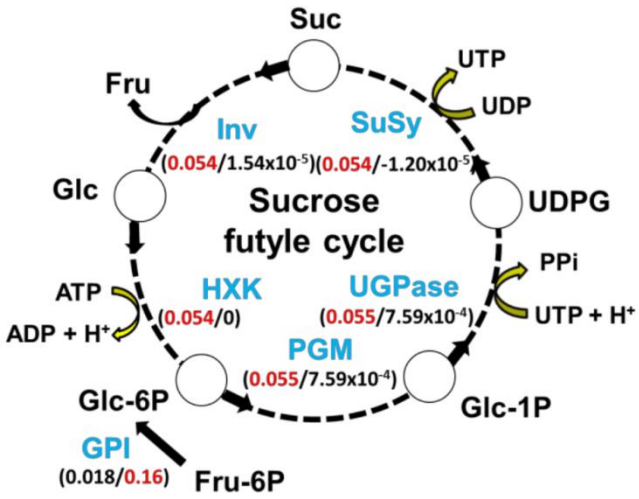


Figure 4.1. (A) A comparison of the predicted metabolic state of G and M cells and (B) detailed depiction of the sucrose futile cycle (SFC) predicted to take place in G cells. *(A) Reactions colored in red (resp. black) carry significantly larger mean flux values in G (resp. M) cells. Reactions depicted in gray cannot be discriminated in terms of mean flux values between the two cell types. The numbers on the reactions correspond to the indices in Supplementary Table S1. The abbreviations used in this figure correspond to: PEP, Phosphoenolpyruvate, Pyr, Pyruvate, Mal, Malate, OAA, Oxaloacetate, Glu, Glutamate, Gln, Glutamine, α – KG, α – Ketoglutarate, FD, Ferredoxin, FD$^-$, reduced*

Ferredoxin, DHAP, Dihydroxyacetone phosphate, G3P, Glyceraldehyde 3-phosphate, G1P, Glucose 1-phosphate, G6P, Glucose 6-phosphate. (B) This cycle is composed by five reactions in which sucrose is preferentially degraded into glucose and fructose by the activity of invertase (Inv, index number 45 in AraCOREred) and resynthesized following activities of hexokinase (HXK, index number 31), phosphoglucomutase (PGM, index number 40), UDP-glucosepyrophosphorylase (UGPase, index number 41) and sucrose synthase (SuSy). In M cells, Glucose 6-phosphate was synthetized exclusively by the action of Glucose 6-phosphate isomerase (GPI, index number 39). Values in parenthesis correspond to the predicted mean flux values for each reaction, values in red correspond to G cells while values in black to M cells. A detailed comparison of the flux values for the reactions in the CBC is provided in Table S3.

In both G and M cells, model simulations predicted a net transport of OAA to the mitochondrion—through three citrate-, isocitrate- and *cis*-aconitate-dependent antiporters— followed by malate production through NAD-MDH and an export of malate back to cytosol (Figure 1A). Further, this pathway was predicted to be the main contributor to the total pool of cytosolic malate in both G and M cells. Specifically, malate was exported out of the mitochondrion with an averaged flux value of 0.311, which constitutes a ~2-fold increase as compared to the mean flux value through the cytosolic NAD-MDH in G cells (Figure 1A, Supplementary Table S1). In the case of M cells, the flux through the malate antiporters, averaging to 0.292, was ~3.5 larger than the cytosolic counterpart, which had a mean flux value of 0.084 (Figure 1A, Supplementary Table S1). Finally, the flux values through the mitochondrial production and export of malate were all significantly larger in G cells (p-value < 1.19×10^{-6}, one-sided Mann-Whitney test, Supplementary Table S1), although the differences were slight—the maximum fold change, ~1.2, corresponded to the mitochondrial NAD-MDH.

Taken together, our predictions suggested that both PEPc – NADP-MDH and the TCA cycle are important contributors to malate synthesis in G cells, although the TCA cycle was the main contributor to the pool. In addition, the marked increment in cytosolic malate production in G cells suggested a diverting pathway to reallocate the excess of cytosolic malate in G cells, especially since mitochondrial malate production was almost the same in the two cell types. This was confirmed by significantly larger flux-sums values of malate in G cells in comparison to M cells, particularly for the case of chloroplasts with a mean flux-sum value in G cells of 2.211 in comparison to 1.493 in M cells (Supplementary Table S2). In fact, model predictions showed a marked 7.5-fold increment in the transport of cytosolic malate to chloroplast in G cells (Figure 1A, Supplementary Table S1).

4.2.3 Chloroplasts adapt their function to meet the metabolic requirements of G cells

Despite decades of research, the role of chloroplasts in G cells and their potential in providing energy for stomatal adjustments or coordination of redox potential is still unresolved. G cells appear to be highly specialized for solute accumulation and are well equipped to generate the energy required for the uptake of ions (*e.g.* K^+, Cl^-),

synthesis of organic anions (particularly malate^{2-}) and accumulation of osmotically active sugars, such as sucrose (Vavasseur & Raghavendra, 2005; Zeiger, Talbott, Frechilla, Srivastava, & Zhu, 2002). Moreover, G cells are known to have fewer and smaller chloroplasts (Willmer & Fricker, 1996) and lower levels of chlorophyll and ribulose-1,5-biphosphate carboxylase/oxygenase (RuBisCO) compared to M cells (Reckmann et al., 1990; Shimazaki, Terada, Tanaka, & Kondo, 1989). Therefore, we hypothesized that G cell chloroplasts are adapted to meet the specific metabolic requirements needed for stomatal functioning, rather than accomplishing the typical tasks of photosynthetic carbon fixation of M cells, *i.e.* to produce sucrose and starch.

To test this hypothesis, we first analyzed the light-dependent reactions in the models specific to G and M cells. We found no significant differences in flux values, neither across the electron transport chain (*i.e.* photosystem II, cytochrome b_6f and photosystem I) nor through the chloroplast ATPase (Figure 1A, Supplementary Table S1). Thus, G cells were predicted to conduct an active photophosphorylation, which was comparable in magnitude to that of M cells. In fact, the flux-sums of ATP in chloroplast were identical in both cell-types (Supplementary Table S2). This result is in agreement with previous observations (Lawson, Oxborough, Morison, & Baker, 2002, 2003), since it has been shown that chloroplasts of G cells are an important source of ATP and NADPH (Azoulay-Shemer et al., 2015; Shimazaki & Zeiger, 1985) and are essential for blue-light induced stomatal opening (Suetsugu et al., 2014).

However, G and M cells produced NADPH differently. Almost all plastidial NADPH was obtained in M cells through the canonical ferredoxin-NADP reductase, which carried a mean flux value of 0.25. In contrast, in G cells, the mean flux value through this reaction was halved (with a flux of 0.126), even though the production of reduced ferredoxin was indistinguishable in both cell types (Figure 1A, Supplementary Table 1). The remaining reduced ferredoxin was predicted to be involved in glutamate synthesis in G cells, through the ferredoxin-dependent glutamate synthase, which carried a mean flux value of 0.124. In contrast, the mean flux value in M cells was only 3.671×10^{-5} (Figure 1A, Supplementary Table S1). The rest of the NADPH in G cells was generated by malate decarboxylation in chloroplasts of G cells by the malic enzyme, to compensate for the lower activity of the ferredoxin-NADP reductase. Interestingly, early studies pointed at malate decarboxylation by malic enzyme as playing a key role in guard cell functioning (Santelia & Lawson, 2016). Moreover, glutamate in chloroplasts of G cells was transported to cytosol in exchange of cytosolic malate by the dicarboxylate transporter. The flux through the latter was predicted to be 7.5-fold larger in G cells, with mean flux values of 0.251 versus 0.033, respectively (Figure 1A, Supplementary Table S1).

Finally, the flux through CO_2 diffusion from chloroplast to cytosol was 17-fold larger in G cells, with mean flux values of 0.230 in G cells and 0.013 in M cells (Figure 1A, Supplementary Table 1). This result largely corresponded to the excess of CO_2 from malate decarboxylation that was not fixed by RuBisCO. Moreover, our predictions

indicated that the exported CO_2 was largely re-fixed by PEPc in the cytosol. These claims can be made since the model incorporates the diffusion of CO_2 to and from the environment, between cellular compartments, as well as the interconversion of CO_2 into bicarbonate (HCO_3). Taken together, these reactions formed a cycle in G cells, where the CO_2 fixed in cytosol was transported as malate to chloroplasts and partly returned to cytosol after malate decarboxylation, with a net production of NADPH (Figure 1A, Supplementary Table S1). These results highlight the adaptation of G cells metabolism to produce malate and NADPH given the lower concentration of chlorophyll and RuBisCO found in these cells (Willmer & Fricker, 1996).

4.2.4 The CBC drives sucrose and starch syntheses in G cells

Several studies have suggested that sucrose acts as an important regulator in G cells and, thus, plays a key role in stomatal movement (Daloso, dos Anjos, et al., 2016). However, the extent to which G cells are able to produce sucrose on their own is a point of debate. On the one hand, due to the low rate of CO_2 fixation, it has been suggested that the contribution of the Calvin-Benson cycle (CBC) to sucrose synthesis in this cell type is negligible. On the other hand, other studies have identified scenarios in which the CBC exhibits a significant activity in G cells (W. H. J. Outlaw, 2003; L. D. Talbott & Zeiger, 1993). Moreover, C fixation by PEPc has been proposed as another route to incorporate C skeletons, which could further be used to drive starch and sucrose synthesis via gluconeogenesis (W. H. Outlaw & Kennedy, 1978; Parvanthi & Raghavendra, 1997; Willmer & Dittrich, 1974). A recent study revealed that G cells can fix CO_2 by both RuBisCO and PEPc (Daloso, Antunes, et al., 2015); however, the extent to which each pathway contributes to the overall amount of sucrose remains an open question.

To resolve the controversy, we investigated the metabolic pathways involved in sucrose synthesis in both G and M cells in our modeling framework. Our results showed that the CBC is active in both cell types. However, most of the reactions involved in the CBC had significantly larger distributions of alternative optimal flux values in M cells, with the notable exception of the PGA kinase, with a mean flux value 1.8-fold larger in G cells (Supplementary Table S3). Our predictions indicated that sucrose synthesis was ultimately driven in both scenarios by the CBC through the canonical pathway of exporting plastidial dihydroxyacetone phosphate (DHAP) and glyceraldehyde 3-phosphate (G3P) to the cytosol, followed by the synthesis of fructose, 1,6–bisphosphate. Interestingly, the model predicted that sucrose synthesis was dominant in G cells, supported by a mean flux value of 0.055 through the sucrose synthase (SuSy) in comparison to 1.5410^{-5} in M cells (Figure 1B). The higher flux through SuSy in G cells was maintained through a futile cycle composed by five reactions (Figure 1B). Futile cycles are metabolic reactions in which the net energy balance or the carbon flux around is zero or near to it (Schwender, Ohlrogge, & Shachar-Hill, 2004), and are known to occur around sucrose in both sink and source tissues (Geigenberger & Stitt, 1991; Trethewey et al., 1998). In our case, sucrose was

re-synthesized from UDP-glucose by activity of SuSy, following the activities of invertase (Inv), hexokinase (HXK), phosphoglucomutase (PGM) and UDP-Glucose pyrophosphorylase (UGPase). These reactions resulted in an equal net consumption and production of UTP and H^+ (Figure 1B). In fact, the marked differences in sucrose flux-sums between G and M cells (mean flux-sum value of 0.11 in G cells in comparison to $5.57x10^{-5}$ in M cells) were due to the contribution of this futile cycle in G cells (Supplementary Table S2).

Although sucrose synthesis and cleavage must be dynamic processes to control stomatal movement, our predictions resulted from invoking the steady-state assumption. Therefore, we interpreted the previously described sucrose futile cycle as the closest steady-state solution to an underlying dynamical process, in which the synthetic and depleting branches alternate in accord with stomatal movements. Given the high number of mitochondria (Willmer & Fricker, 1996), the large catabolic activity found in G cells (Hampp, Outlaw, & Tarczynski, 1982) and the importance of osmolyte accumulation in this cell type (Zhu, Talbott, Jin, & Zeiger, 1998), the identified futile cycle could represent a mechanism that allows avoidance of excess starch synthesis. As a result, carbon skeletons are maintained in sucrose and hexoses, rather than starch, and can be readily used as substrate for glycolysis and mitochondrial metabolism.

Given that the CBC was predicted to be active in G cells, we investigated whether it also drives starch synthesis. We found that G cells were predicted to conduct starch synthesis, although fluxes were significantly higher in M cells in two of the total three reactions involved (Supplementary Table S3). Mean flux values throughout starch degradation were in general small in both cell-types, although this process was significantly pronounced in M cells. For instance, mean flux values through the amylase were ~8-fold larger in M cells, and the disproportionating enzyme was predicted to be active only in M cells (Supplementary Table S3). These results suggested that starch degradation was not a major player in sucrose synthesis in G cells.

We also found marked differences between G and M cells in the main source of CO_2 entering the CBC. Cytosolic CO_2 diffusion to chloroplast was only present in M cells (mean flux value of 0.01, Figure 1A, Supplementary Table S1). Conversely, cytosolic malate import to chloroplast by the dicarboxylate transporter, followed by decarboxylation by plastidial MDH, was the main source of CO_2 in G cells (Figure 1A, Supplementary Table S1). In addition, as commented in the previous section, cytosolic PEPc was key to driving malate synthesis in G cells and its import to chloroplast. Altogether, these results match the experimental observations from (Daloso, Antunes, et al., 2015) and suggest that both, carbon fixation through CBC and cytosolic PEPc followed by gluconeogenesis play a major role in driving sucrose synthesis in G cells. As a result, the findings in (Daloso, Antunes, et al., 2015) serve as validation of our approach to integrating transcriptomics data for the purpose of comparing the distribution of values for particular fluxes at alternative optima.

Moreover, they indicate the presence of a C_4-like metabolism in G cells, in which the CO_2 fixation by RuBisCO is derived from decarboxylation of the C_4 acid malate.

4.2.5 Robustness of prediction to adding constraints derived from experimental observations

Our computational results presented were generated by constraining the fluxes with G cell- and M-specific expression data. Therefore, no assumptions about the activity of particular reactions were considered—besides imposing a minimal flux through biomass production and the energy maintenance reactions—as to avoid biased predictions. However, we observed two modeling predictions that were unlikely under the photosynthetically active scenario evaluated here. In the first case, RuBisCO oxygenation was absent in both G and M cells, while experimental evidence constraints the ratio of RuBisCO's carboxylation to oxygenation to be within 1.5 and 4 for both cell types (F. Ma, Jazmin, Young, & Allen, 2014; Sharkey, 1988; Szecowka et al., 2013). In the second case, three reactions in the CBC: Fructose,1,6-Bisphosphatase, sedoheptulose 1,7-bisphosphate aldolase and sedoheptulose-1,7-bisphosphatase (reaction numbers 11, 13 and 14 in AraCORERed) carried very low or no flux values, thus compromising the functional integrity of the CBC.

To address these inconsistencies, we repeated the computational analysis adding this time a constraint on the carboxylation to oxygenation ratio and including a minimum flux value through the three mentioned reactions in the CBC (Materials & Methods). We next evaluated the qualitative changes upon the inclusion of these additional constraints on the main computational results previously generated. To this end, we looked at the differences in the outcomes of the Mann-Whitney tests—comparing the fluxes through each reaction in G and M cells—between the modeling predictions when no additional constraint was considered and upon the inclusion of the carboxylation to oxygenation and the minimum flux value constraints discussed above. We found that 26.67% of the reactions in AraCORERed changed the Mann-Whitney test status when including the carboxylation to oxygenation constraint and the minimum flux value constraints through the CBC (this figure was reduced to 26.1% when reactions in the CBC, directly affected by the imposed constraints, were not taken into account). However, the vast majority of these changes did not qualitatively affect the main results presented in this study. For instance, the G/M mean flux ratio through the triad CA, PEPc and cytosolic NADP-MDH, shifted from ~12, ~12 and ~2 to ~37, ~37 and ~2, and the mean flux ratios through the ferredoxin NADP-reductase and the glutamate synthase shifted from 0.502 to 0.520 and from ~3391 to ~719, respectively (Supplementary Tables S6, S7 and S8 display the full list for comparison). Therefore, our main results are robust upon the inclusion of these additional, biologically relevant, constraints.

4.2.6 Validation of model predictions

In this section, we provide a description of the findings from an independent gas chromatography mass spectrometry (GCMS)-based ^{13}C-labelling experiment which we employed to validate the flux-based predictions. The employed GCMS approach does not allow us to analyse the ^{13}C flux distribution in intermediates from CBC and glycolysis. Therefore, we focused the analysis on sucrose and metabolites related to photorespiration, amino acid metabolism, anaplerotic CO_2 fixation and the TCA cycle.

4.2.7 G cells have higher anaplerotic CO_2 fixation

The anaplerotic reaction catalysed by PEPc is characterized by the incorporation of a molecule of HCO_3 into PEP producing OAA and Pi (Melzer & O'leary, 1987). The C fixed by PEPc is incorporated in the fourth C of OAA (Nargund, Misra, Zhang, Coleman, & Sriram, 2014), which can be directly converted to Asp or malate by aspartate amino transferase (AspAT) or MDH, respectively. The anaplerotic CO_2 fixation is the main source of C incorporation in cells with C_4 or CAM metabolism, in contrast to C_3 cells (Jiao & Chollet, 1991; Osmond, 1978). It has been hypothesized that the anaplerotic CO_2 fixation by PEPc activity is higher in G cells in comparison to M cells (Daloso, Antunes, et al., 2015; Reckmann et al., 1990; Vavasseur & Raghavendra, 2005). This idea is supported by the higher expression of genes related to this pathway in G cells in comparison to M cells (Bates et al., 2012; Bauer et al., 2013; Leonhardt et al., 2004; R.-S. Wang et al., 2011). Recent results from a ^{13}C-isotope labelling study strongly suggest that G cells are able to fix CO_2 by both pathways those catalysed by RuBisCO and PEPc (Daloso, Antunes, et al., 2015). However, despite the evidences pointing for a differential anaplerotic activity in G cells, this hypothesis has not yet been adequately tested.

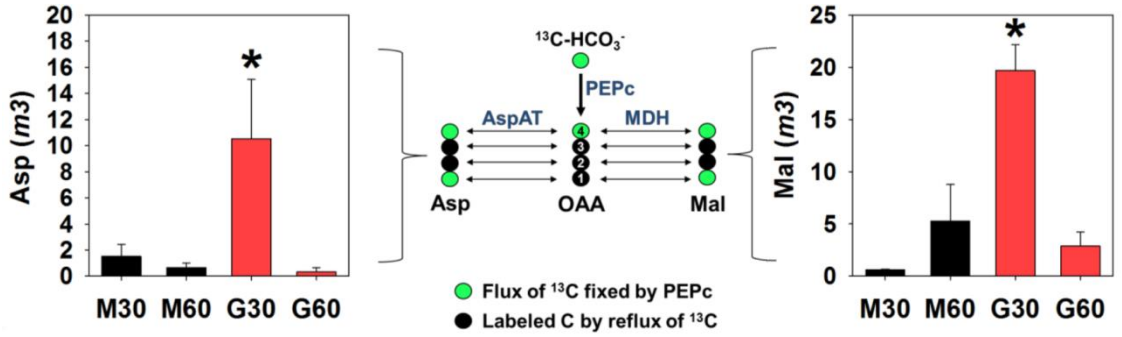


Figure 4.2. Evidence for the higher anaplerotic CO_2 fixation in G cells in comparison to M cells. *M cells (black bars) and G cells (red bars) were fed with ^{13}-$NaHCO_3$ and harvested after 30 min and 60 min in the light. The abundance of mass isotopomers of aspartate m3 (left side) and malate m3 (right side) in mesophyll cells (M) or guard cells (G) after 30 and 60 min in the light is displayed. The anaplerotic reaction catalysed by phosphoenolpyruvate carboxylase (PEPc) and the subsequent steps catalysed by aspartate aminotransferase (AspAT) and malate dehydrogenase (MDH) are highlighted in the center of the figure. Small spheres represent carbon atoms labelled directly by the activity of PEPc (green spheres) or by the reflux of this ^{13}C by the activity of the tricarboxylic acid cycle (black spheres). Asterisks indicate values that are significantly different between mesophyll and guard cells by Student's t-test (P <0.05) in the same time point. Data presented are mean ± standard deviation (n = 3).*

Here, we used ^{13}C-isotope labelling approach to validate the predictions about anaplerotic CO_2 fixation. Given the instability of OAA in GCMS-based analysis, we focused on the malate and Asp given that these metabolites are primary products of OAA conversion. The relative isotopomer analysis revealed that the full labelled ion (*m3*) of malate and Asp was ~34- and ~7-fold higher in G than M cells after 30 min under light, respectively (Figure 2). The level of the fully labelled ion of Asp and malate decreased from 30 to 60 min under light, suggesting that these metabolites are degraded or exported out of G cells. Although we cannot exclude the possibility of efflux of these metabolites from G cells, they can also be employed to increase the flux through the TCA cycle. This idea is supported by the increase in the ^{13}C-enrichment in citramalate, succinate and (to a lower extent) fumarate after 60 min under light (Figure 3).

The higher level of Asp and malate *m3* led to a higher ^{13}C-enrichment in these metabolites in G cells in comparison to M cells (Figure 3). In analyses that take into account the concentration of the metabolites, we also found higher percentage (%) and total ^{13}C-enrichment in Asp and malate in G cells (Tables S9 and S10). The fully labelled malate is not only due the PEPc activity, but it also depends on labelled C from glycolysis and the TCA cycle. As stated above, PEPc fixes CO_2 onto the fourth C of OAA, which can be then converted to malate, producing malate with maximum of two ^{13}C (refer to green spheres on Figure 2). Therefore, the other ^{13}C detected in malate and Asp obligatorily comes from fully labelled Acetyl-CoA, which is derived from glycolysis and its assimilation provides two additional ^{13}C to metabolites of, or

associated to, the TCA cycle (Daloso, Antunes, et al., 2015). These results were in line with the predictions about larger flux-sums of malate in G in comparison to M cells (Supplementary Table S2). Further, G cells showed higher ^{13}C-enrichment in metabolites that can be derived from Asp (*e.g.* homoserine and threonine) and malate (*e.g.* succinate and citramalate) (Figure 3, Supplementary Table S11). Altogether, these results confirmed the modelling predictions and reveal that the anaplerotic CO_2 fixation catalysed by PEPc is higher in G cells.

4.2.8 Guard cells have higher ^{13}C-enrichment but lower capacity to produce sucrose under ^{13}C-$NaHCO^-_3$

Sucrose is the main metabolite translocated throughout the plant and performs several functions in the metabolism (Fettke & Fernie, 2015). Sucrose can be produced by using triose phosphates and hexoses exported from chloroplast following photosynthesis and starch degradation, respectively (Lunn, 2008), as well as by PEPc fixation and gluconeogenesis (Eastmond et al., 2015). The capacity of G cells to produce sufficient quantity of sucrose has long been debated. Under the experimental condition used here, we observed higher ^{13}C-enrichment in sucrose after ^{13}C-$NaHCO^-_3$ incorporation in G cells (Figure 3), which confirms the predictions of the model (Figure 1A-B). However, it is important to note that M cells had, on average, 2.2-fold more sucrose than G cells (Supplementary Table S12). The difference in the amounts of sucrose between the two cell types leads to an equal percentage of ^{13}C-enrichment and total ^{13}C-enrichment in sucrose between G and M cells, with both calculations cosidering the amount of the metabolite of each cell type (Tables S9 and S10). Moreover, the higher ^{13}C-enrichment observed in G cells (Figure 3) may be due the use of HCO_3 as labelled substrate, which could favour the fixation by PEPc rather than by RuBisCO. The higher expression of CA (Aubry et al., 2016; Leonhardt et al., 2004) and higher malate decarboxylation in the chloroplast of G cells, as predicted by the model, may create a high CO_2-concentrated atmosphere around RuBisCO of G cells, similarly to what has been observed in C_4 cells. This would optimize the plastidial CO_2 fixation by RuBisCO in these cells, leading to higher ^{13}C-enrichment in sucrose. This idea is further supported by the higher ^{13}C-enrichment observed in metabolites from photorespiratory pathway such as Ser and glycerate (Figure 3, Supplementary Table S11). However, further experimental evidence is needed to confirm this hypothesis and the model predictions.

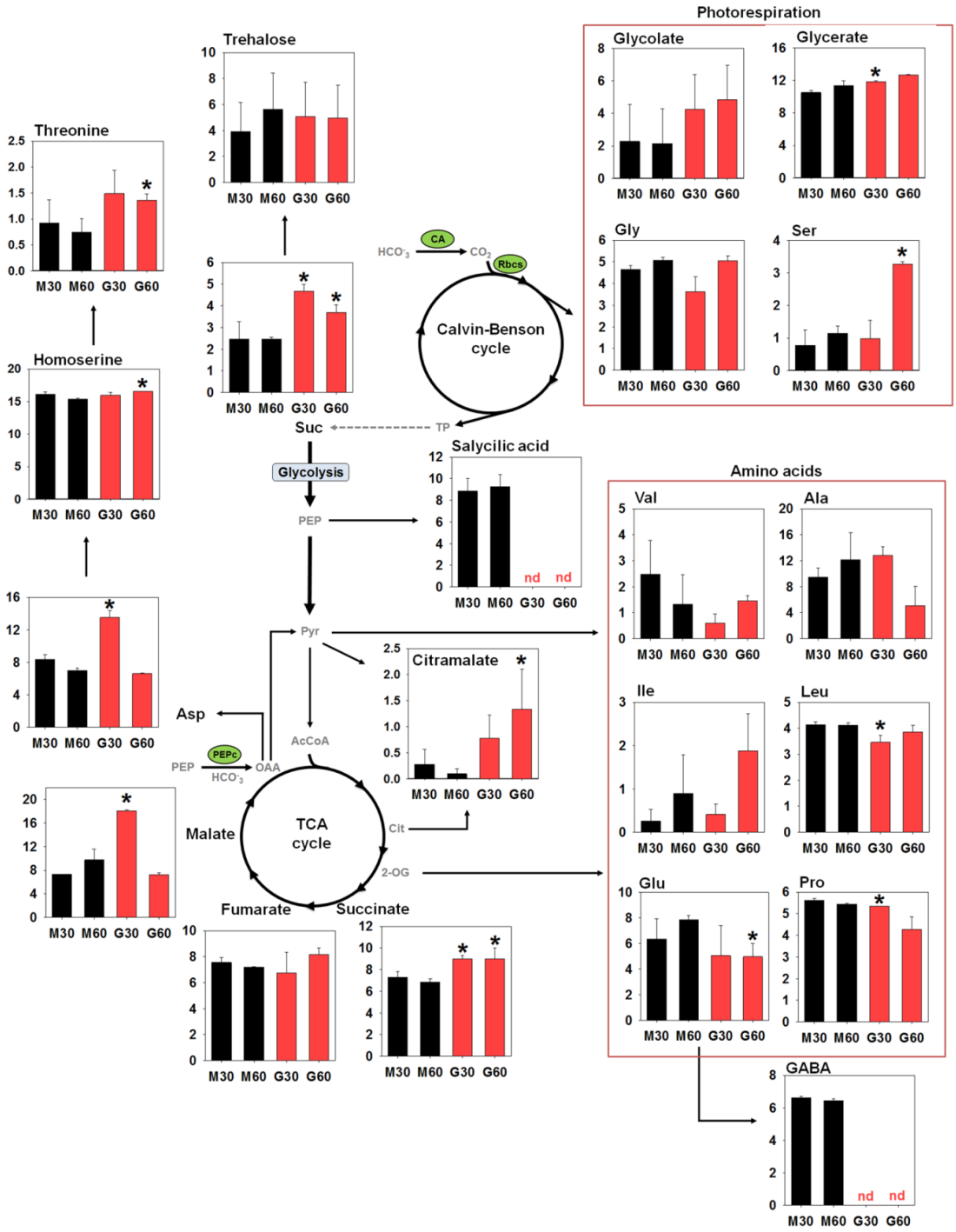


Figure 4.3. ***^{13}C-enrichment in primary metabolites.*** *M cells (black bars) and G cells (red bars) were fed with ^{13}-$NaHCO_3$ and harvested after 30 min and 60 min in the light. Asterisks indicate values that are significantly different between M and G cells, (Student's t-test, P <0.05) for the same time point. Data presented are mean ± standard deviation (n = 3). The complete list of the ^{13}C-enrichment is presented in Supplementary Table S11. Abbreviations: metabolites: GABA, gamma-aminobutyric acid; Suc, sucrose. Enzymes: CA, carbonic anhydrase; PEPc, phosphoenolpyruvate carboxylase; Rbcs, ribulose-1,5-biphosphhate carboxylase/oxygenase. Amino acids are abbreviated using the standard three-letters code.*

Sucrose was thought to act as an osmolyte for G cell regulation (L. Talbott & Zeiger, 1998). However, recent evidences suggest that the role of sucrose for G cells regulation may be primarily energetic (Antunes, Provart, Williams, & Loureiro, 2012; Daloso, Antunes, et al., 2015; Daloso, Williams, et al., 2016; Ni, 2012) and that sugar metabolism and HXK activity may be pivotal in the control of stomatal movements (Kelly et al., 2013; Lugassi et al., 2015). The model used here predicted that higher fluxes through sucrose occur in G cells and this can be explained by a substrate (futile) cycle formed around this metabolite. Substrate cycles have been proposed to be important for the regulation of plant metabolism (Alonso, Raymond, Rolin, & Dieuaide-Noubhani, 2007; Geigenberger et al., 1997; Geigenberger & Stitt, 1991; Hargreaves & ap Rees, 1988; Hill & ap Rees, 1993) (Alonso, Vigeolas, Raymond, Rolin, & Dieuaide-noubhani, 2005), despite their ‘‘futile’’ designation. It is known that sucrose and hexose cycles are supported by high activity of enzymes such as SuSy, Inv, HXK, sucrose-phosphate synthase (SPS) and other sugar-related enzymes (Alonso et al., 2007; Dancer, Hatzfeld, & Stitt, 1990; Wendler, Veith, Dancer, Stitt, & Komor, 1991). Interestingly, most of these enzymes are highly expressed in G cells (Daloso, dos Anjos, et al., 2016), in further support for the idea of the substrate cycle predicted by the model. Additionally, futile cycles have been confirmed *in vivo* by steady-state and pulse-labelling approaches using both ^{14}C and ^{13}C substrates (Alonso et al., 2005), which can be used and are required to confirm our model predictions.

4.3 Conclusions

Despite decades of research, the role of central carbon metabolism on the functions of G cells remains poorly understood. Here, we used transcriptomics data and a large-scale metabolic model to predict pathways with differential flux profiles between G and M cells. Our analysis pinpointed reactions whose distributions of fluxes in the space of alternative optima differ between G and M cells. Since reaction fluxes are difficult to be experimentally estimated in photoautotrophic growth conditions, we predicted flux-sums as descriptors of metabolite turnover and validated the qualitative behavior via an independent ^{13}C-labeling experiment. Our results highlighted the metabolic differentiation of G cells as compared to the surrounding M cells, and strengthen the idea of occurrence of a C_4-like metabolism in G cell, as evidenced by the higher anaplerotic CO_2 fixation in this cell. Moreover, our modeling approach brings important and new information concerning CBC and sucrose metabolism in G cells, indicating that the main source of CO_2 for RuBisCO comes from malate decarboxylation rather than CO_2 diffusion and that G cells have a futile cycle around sucrose. The modeling and data integration strategy can be used in future studies to investigate the concordance between flux estimates with data from different cellular layers. In addition, future studies on guard cell physiology would benefit from coupling the flux-centered genome-scale modeling framework presented in this study with existing kinetic models of stomatal movement, such as OnGuard (Hills et al., 2012). Finally, although still technically challenging, future studies would also benefit

from quantitative experimental data of coupled G and M cells *in vivo*, which could be integrated in a unified modeling framework addressing the coordination between the two cell types.

4.4 Material and methods

This section provides the details of the computational methods used in the metabolic modeling. A depiction of the general procedure followed is also available in Figure S4.1. In addition, all the MATLAB code used to obtain the predictions is provided in File S1.

4.4.1 Gene expression data

G cell gene expression data was obtained from (Bates et al., 2012) published under the GEO accession numbers GSM918075, GSM918076 and GSM918077, which correspond to three replicates of the same experiment. Expression data of M cells were obtained from (Pandey et al., 2010), also with three replicates with accession numbers GSM486916, GSM486917, and GSM486918. In both cases, expression data were measured from wild type Col-0 *A. thaliana*. Data from G cells set was obtained from plants were grown for 8-10 weeks at 22 °C, and in an 8 hours light/16 hours dark cycle under 150 $\mu mol.m^{-2}.s^{-1}$. In the case of the data set from M cells, plants were grown for 5 weeks at 20/16 °C, and in an 8 hours light/16 hours dark cycle under 120 $\mu mol.m^{-2}.s^{-1}$. CEL files were normalized using the RMA method implemented in the *affy* R package (Gautier et al., 2004). In addition, probe names were mapped to gene names following the workflow described in (Moyano et al., 2015), where probes mapping to more than one gene name are eliminated.

Expression values were mapped to reactions following the gene-protein-reaction rules and a self-developed MATLAB function, *mapgene2rxn*, which can be found in File S4.1. Specifically, the conditional relation g_i AND g_j in a given reaction rule was modeled as the minimum expression value of the two genes, g_i, g_j. The conditional relation g_i OR g_j was modeled as the maximum expression of the two genes. This process was repeated for each of the three replicates in each cell-type (*i.e.*, G and M cells). The mean and standard deviation among replicates were then calculated for each reaction associated gene expression. Finally, values were scaled to the maximum in each experiment to obtain the final expression data used in this study.

4.4.2 Metabolic network model

AraCORE, a metabolic network model of the primary metabolism of *A. thaliana* developed by (Arnold & Nikoloski, 2014) was used to reconstruct the metabolic networks specific to G and M cells. The model includes 549 reactions and 407 metabolites assigned to four subcellular compartments. The original AraCORE

contains exchange reactions that directly link organelles to the environment (*i.e.* circumventing the cytosolic compartment). Therefore, all exchange reactions bypassing the cytosol were removed to avoid biased results. Therefore, here, we used a reduced AraCORE version (AraCORERed), available in File S1 that consists of 455 reactions and 374 metabolites.

4.4.3 Gene expression integration in AraCORERed

The two context-specific flux distributions (*i.e.,* corresponding to guard cell and mesophyll) were obtained by integrating the expression data into AraCORERed. To this end, we used the $\text{RegrEx}_{\text{LAD}}$ method (Chapter 3, section 3.2.1.1), which performs an optimization process to find a feasible flux distribution (*i.e.*, satisfying the stoichiometric and thermodynamic constraints dictated by the metabolic model used) that maximizes the concordance to the integrated expression data. In addition, for each cell type, the mapped data were scaled by the respective standard deviations (estimated from the three available replicates. In this way, reactions whose associated data were less consistent among replicates (as quantified by the standard deviation) contributed to a lesser extent to the global similarity between the optimal flux distribution and the integrated expression data.

The biomass reaction was forced to be active by imposing that the flux, v_{bio}, through the reaction "*light-dependent biomass*" (number 454 in AraCORERed) satisfies the constraint $v_{bio} \geq 10^{-6}$. Additionally, the fluxes through the three energy maintenance reactions in AraCORERed were forced to be greater or equal to 0.001. The three maintenance reactions represent the consumption of ATP by non-metabolic processes—*i.e.,* apart from the consumption in the reactions included in AraCORERed—in the cytosol (reaction index number 448), chloroplast (449) and mitochondrion (450), respectively. The lower bound values were chosen to represent approximately a 20% of the theoretically maximum flux value for each reaction in the alternative optima space of $\text{RegrEx}_{\text{LAD}}$ (as calculated per $\text{RegrEx}_{\text{FVA}}$, described in the next section).

4.4.4 Evaluation of the alternative optima space

We investigated the space of alternative $\text{RegrEx}_{\text{LAD}}$ solutions with two complementary approaches. We employed $\text{RegrEx}_{\text{AOS}}$ (described in section 3.2.1.1) to generate a uniform sample $V_{AO} = \{v^*_{i,k}, i = 1, \dots, N_R, k = 1, \dots, n\}$, of $n = 2 \cdot 10^4$ random alternative optimal flux distributions (containing $N_R = 455$ reactions) for each of the cell-specific scenarios. On the other hand, we developed and used $\text{RegrEx}_{\text{FVA}}$, a FVA-like (for Flux Variability Analysis) algorithm, to compute the minimum and maximum allowable flux values in the alternative optima space. $\text{RegrEx}_{\text{FVA}}$, depicted in (4.1), adapts the Flux Variability Analysis(R. Mahadevan & Schilling, 2003) procedure—originally designed to investigate the alternative optima

space of the linear programs behind Flux Balance Analysis (Orth et al., 2010)—to the particular computational setup of $\text{RegrEx}_{\text{LAD}}$.

$$
\begin{aligned}
&\min/\max_{\substack{\epsilon^+=[\epsilon^+_{irr};\epsilon^+_{for};\epsilon^+_{back}],\\ \epsilon^-=[\epsilon^-_{irr};\epsilon^-_{for};\epsilon^-_{back}],\\ v=[v_{irr};v_{for};v_{back}]\in\mathbb{R}^+_0,\\ x\in\{0,1\}^n}} \quad v_j \\
&s.t. \\
&1-11 \quad (3.1) \\
&12.\ w^T(\epsilon^+ + \epsilon^-) = w^T(\epsilon^+_{opt} + \epsilon^-_{opt}) \\
&13.\ \|v\|_1 = \|v_{opt}\|_1 .
\end{aligned}
\tag{4.1}
$$

In (4.1), constraints 1-11 are inherited from (3.1) the MILP followed by the $\text{RegrEx}_{\text{LAD}}$ algorithm of section 3.2.1.1. Constraints 12 and 13 are identical to the homonymous constraints in (3.2) the MILP followed by $\text{RegrEx}_{\text{AOS}}$ (section 3.2.1.1). Supplementary Tables S4 and S5 show the extreme values for the reactions displayed in Supplementary Tables S4.1 and S4.3 (a complete list can be found in the MATLAB data file in Supplementary File S1).

4.4.5 Evaluation of flux values across the alternative optima space

Next, the previously generated distributions of alternative flux values of each reaction were compared between G and M cells. To this end, a Mann–Whitney test (Lovric, 2011) (*ranksum* MATLAB function) was applied to obtain the set of reactions showing significantly increased flux values across the alternative optima space for each cell-type. Specifically, we performed a right-tailed test with null hypothesis stating that there were not differences between the two cell types and alternative hypothesis stating that one cell-type (*i.e.,* guard cells or mesophyll depending on the comparison) had a bigger flux distribution than the other one, rejecting the null hypothesis at the significance level of $\alpha = 0.05$. In addition, we performed a two-tailed Mann-Whitney test evaluating only the significance of the difference between two distributions, *i.e.*, with null hypothesis stating no differences and alternative hypothesis stating significant differences (either larger or smaller) between the two distributions. In this study, only distributions that passed the two tests, *i.e.*, significantly larger (or smaller) and significantly different, were taken into account, as to prevent inconsistent results.

4.4.6 Evaluation of flux-sum values across the alternative optima space

We calculated the flux-sum values for each metabolite m ($f_{m,k}$) in the AraCOREred model and for each alternative optimal flux distribution, $v_k^* \in V_{AO}$ and cell-type as follows:

$$f_{m,k} = \sum_j |v_{j,k}^*|, j \in R_m \quad (4.2)$$

where R_m is the index set corresponding to reactions in which metabolite m participates either as a substrate or as a product. This procedure generated a distribution of alternative flux-sum values for each metabolite in each cell-type. Next, the previously generated distributions of flux-sum values of each metabolite were compared between G and M cells. To this end, we applied the same battery of Mann-Whitney tests previously used to compare the distributions of alternative optimal flux values.

In this analysis, the different subcellular localizations of a given metabolite were treated as different metabolites in the metabolic network, due to the compartmentalization of the AraCOREred model—which is subdivided into cytosol, mitochondrion, chloroplast, and peroxisome. Therefore, the distribution of flux-sum values, $f_{m,k}$, presented above, was obtained specifically for each subcellular localization of a given metabolite. However, the metabolic experimental data generated in this study do not discriminate between subcellular localizations of the measured metabolites—*i.e.*, the data measure the total cellular pool of a metabolite and not the specific concentrations in each subcellular compartment. To match the experimental conditions, we additionally calculated the flux-sums of the metabolites with experimental data across all subcellular compartments. Specifically, in (1), the reaction index set, R_{m*}, of a metabolite, m^*, with experimental data, contained all reactions in which m^* participated across all subcellular compartments. The same statistical analysis used to compare the flux-sum distributions of compartmentalized metabolites was applied in this case.

4.4.7 Integration of additional constraints derived from experimental observations

Bounds on the carboxylation to oxygenation ratio of RuBisCO were included in the following way. Let v_C denote the flux through RuBisCO carboxylation and v_O that of the oxygenation, then the non-linear constraint

$$r_{lb} \leq \frac{v_c}{v_o} \leq r_{ub} \quad (4.3)$$

can be transformed into the pair of linear constraints

$$r_{lb}v_o - v_c \leq 0, \tag{4.4}$$

$$r_{ub}v_o - v_c \geq 0, \tag{4.5}$$

where r_{lb}, r_{ub} respectively denote the lower and upper bound of the carboxylation to oxygenation ratio. These linear constraints were integrated as additional constraints to the optimization programs performed by RegrEx$_{LAD}$ and RegrEx$_{AOS}$ (presented in section 3.2.1 of Chapter 3) to guarantee that any alternative optimal solution agreed with the specified bounds (in this study $r_{lb} = 1.5, r_{ub} = 4$). Constraints regarding the minimum flux through the reactions: *Fructose,1,6-Bisphosphatase*, *sedoheptulose 1,7-bisphosphate aldolase* and *sedoheptulose-1,7-bisphosphatase* (reaction number 11, 13 and 14 in AraCOREred) where integrated by increasing the lower bound through these reactions from zero to a small amount (0.001 in this study).

To evaluate the changes in the simulation results due to the integration of the new set of constraints, we compared the outcomes of the Mann-Whitney tests across all reactions in the AraCOREred model. Concretely, we first transformed the vector of p-values resulting from the comparison between G and M cells of the distributions of alternative optimal flux values for each reaction, into a binary vector. To this end, p-values below the significance threshold $\alpha = 0.05$ were mapped to 0, and the rest to 1. This process was repeated for each scenario: i) the original results without additional constraints, ii) the results generated after constraining the carboxylation to oxygenation ratio and iii) the results generated when constraining the carboxylation to oxygenation ratio and the flux through the three above mentioned reactions. We next computed the Hamming distance between the three binary vectors. In this case, we evaluated the distance between the whole set of reactions in the AraCOREred model, and between all reactions except those from the CBC, since reactions in the CBC were directly affected by the newly imposed constraints.

4.4.8 Plant material and growth conditions

Seeds of wild type *Arabidopsis thaliana* L. plants (Columbia ecotype) were handled as described previously (Daloso, Müller, et al., 2015). Fully expanded rosette leaves of 5-week-old plants grown under long day conditions (16h light/8h dark), light intensity 100 µmol photons m^{-2} s^{-1} and temperature 20°C ± 2 were harvested for isolation of both G cells and mesophyll cell protoplasts (MCP).

4.4.9 Experimental set-up for *in vivo* G and M cells analyses

We recently developed a methodology to perform ^{13}C kinetic isotope labeling experiments in isolated G cell enriched epidermal fragments (Daloso, Antunes, et al., 2015). Here, we modified this method to analyze the metabolic flux distribution in

simultaneously isolated G cells and MCP. Several experiments were performed to simultaneously isolate both cell types from the same plant material as well as to perform a ^{13}C kinetic isotope labeling experiment following the metabolic fate of ^{13}C-$NaHCO_3$ by gas chromatography-time of flight-mass spectrometry. All the solutions used for G cells and MCP isolation were prepared in deionized water and filtered through a 0.45 µm filter. The isolation of both cell types was carried out in the dark using leaves from dark-adapted plants in order to avoid light induced metabolic changes during the isolation of both cell types. Furthermore, in contrast to the original protocols in which the isolated cells are subjected to a high concentration of mannitol (0.4-0.56 M), we decided to reduce the mannitol concentration to minimize the excess of this metabolite in the final steps of the protocol, since this causes problems in subsequent metabolite determination Thus, the concentration of this osmolyte in the medium was reduced gradually from 0.4 M (solution I) to a final concentration of 0.05 M (solution IV - see below). The solutions used for GC and MCP isolation were: **enzymatic solution -**20 mMMes/NaOH, pH 5.7, 0.4 M mannitol, 10 mM $CaCl_2$, 20 mMKCl, 0.1% (w/v) bovine serum albumin (BSA), 1% (w/v) cellulase, Onozuka R10 (Yakult Pharmaceutical Industry Co., Tokyo, Japan), 0.25% (w/v) macerozyme, Onozuka R10 (Yakult Pharmaceutical Industry Co., Tokyo, Japan). **Solution I -** 0.4 M mannitol, 1 mM $CaCl_2$.**Solution II -** 20 mMMes/NaOH, pH 6.5, 0.1 M mannitol, 1 mM $CaCl_2$. **Solution III -** 20 mMMes/NaOH, pH 6.5, 0.05 M mannitol, 1 mM $CaCl_2$, 5 mMKCl and **Solution IV -** 20 mMMes/NaOH, pH 6.5, 0.05 M mannitol, 1 mM $CaCl_2$, 5 mMKCl, 1 m$M^{13}C$-$NaHCO_3$.

4.4.10 ^{13}C isotope labelling experiment using isolated M cell protoplasts

Arabidopsis MCP were isolated from dark adapted five-week-old plants using the TAPE-sandwich method (Wu et al., 2009) with modifications. Approximately 20 leaves per replicate were peeled and placed in a petri dish containing 50 mL of enzymatic solution and shaken in the dark. After 90 min, the solution containing MCP was transferred to a 50 mL Falcon tube and centrifuged at 100 *g* for 15 min at 4°C. The supernatant was removed and the pellet containing MCP was gently re-suspended in solution I and kept on ice and in the dark for 30 min. This procedure was repeated by adding and removing the solutions II and III with the same interval on ice and in the dark. After the addition of the solution IV, the MCP were immediately transferred to the light (45 ± 1 µmol photons m^{-2} s^{-1}) and harvested after 30 and 60 min. We adjusted a methodology to rapidly collect and frozen MCP in the light following a previous established methodology developed for ^{13}C kinetic labelling experiments in Algae (Krall, Huege, Catchpole, Steinhauser, & Willmitzer, 2009) which the MCPs were vacuum concentrated to a glass filter (1.6 µm). This process was carried out under the same light source used. The time spent between the transfer of the MCP from the petri dishes to the glass filter and the subsequently frozen was around 1-2 min.

4.4.11 ^{13}C kinetic isotope labelling experiment in G cells

Arabidopsis G cells were isolated from five-week-old plants according to previous methods (Daloso, Antunes, et al., 2015; Misra, De Armas, Tong, & Chen, 2015) with minor modifications. Approximately 30 whole rosettes were ground using a commercial blender with an internal filter (Philips, HR 2084, Amsterdam, The Netherlands) containing 300 ml of cold deionized water for 3 min. The internal filter is important to remove the excess of fibers and mesophyll cells (Daloso, Antunes, et al., 2015). After that, the isolated guard cell enriched epidermal fragments were collected on a 220 μm nylon mesh and rinsed well with distilled water (1.5 L). After drying the excess of water, the G cell enriched epidermal fragments preparation was transferred to the enzymatic solution and kept for 90 min in the dark with shaking (Misra, De Armas, et al., 2015). The guard cell enriched epidermal fragments were collected on a 30 μm nylon mesh, rinsed with solution I and kept in 15 mL of this solution for 30 min on ice and in the dark. The osmotic potential of the solution was decreased by adding 15 mL of the solution II and III with an interval of 15 min on ice. After, G cell enriched epidermal fragments were collected on a 30 μm nylon mesh and transferred to 5 mL of solution III and carefully layered on top of 20 mL Histopaque® solution (Histopaque-1077, Sigma Aldrich, St. Louis, USA) in a 50 mL falcon tube. The tube was centrifuged at 200 *g* for 15 min in order to separate GCs from trichomes and other cell debris. The layer of G cells was withdrawn from the interface of the two solutions with a 5 mL pipette, collected on a 30 μm nylon mesh, transferred to a falcon tube containing solution IV and transferred to the light (45 ± 1 μmol m^{-2} s^{-1}). After 30 and 60 min under light, G cells were rapidly vacuum concentrated to a glass filter (1.6 μm) as performed for MCP and frozen.

4.4.12 Extraction and analysis of metabolites

The extraction of polar metabolites from G cells and MCP were carried out following an established gas chromatography-time of flight-mass spectrometry based platform (Lisec, Schauer, Kopka, Willmitzer, & Fernie, 2006) adapted to G cells (Daloso, Antunes, et al., 2015). In brief, the extraction of the metabolites was carried out using 1000 μL of methanol (100%) at 70 ºC for 1 h with constant agitation. 60 μl of Ribitol (0.2 mg/ml stock in dH_2O) was added as an internal quantitative standard. The extract was centrifuged at 11000*g* for 10 min, and 600 μL of the supernatant was transferred to another tube, where 500 μL of chloroform (100%) (LC grade, Sigma) and 800 μL of deionized water were added. After vortexing for 10 s, another centrifugation was carried out for 15 min at 2200 *g*. 1300 μL of the (upper) polar phase was collected, transferred to 2.0 ml tubes, and reduced to dryness in a speed vac. The sample derivatization was carried out using *N*-Methyl-*N*-(trimethylsilyl) trifluoroacetamide (MSTFA, CAS 24589-78-4, Macherey& Nagel, Düren, Germany) and methoxyamine hydrochloride (CAS 593-56-6, Sigma, Munich, Germany) dissolved at 20 mg/ mL in pure pyridine (CAS 110-86-1, Merck, Darmstadt, Germany) (Lisec *et al*., 2006).

Metabolites were identified by comparison the Golm metabolome database (Kopka et al., 2005). The analysis of relative abundance of mass isotopomers was carried out using Xcalibur 2.1 software (Thermo Fisher Scientific, Waltham, MA, USA) exactly as described in ref(Daloso, Antunes, et al., 2015). Absolute levels and percentage and total ^{13}C-enrichment of metabolite was determined as described previously (Huege, Goetze, Dethloff, Junker, & Kopka, 2014; Roessner et al., 2001).

Acknowledgements

SRE would like to thank the Max Planck Society and the International Max Planck Research School “Primary Metabolism and Plant Growth” for providing the funding.

Author Contributions

Z.N. designed the study. S.R-E. performed the computational analyses and interpreted the finding. D.M.D. and A.R.F. designed the experimental aspects of the study. D.M.D. and Y.Z. performed the experiments. D.M.D. analyzed the labeling data. All authors contributed writing the manuscript.

Chapter 5

Discussion

The scientific contributions of this thesis have been threefold. Chapter 2 presented and tested the performance of RegrEx, a new method to integrate experimental data into GEMs and aimed at generating context-specific flux distributions and models. Chapter 3 dealt with alternative optima in context-specific metabolic predictions, introduced two new computational methods to analyze the alternative optimal space of RegrEx: $\text{RegrEx}_{\text{FVA}}$ and $\text{RegrEx}_{\text{AOS}}$, and extended this analysis to MBA-like methods (defined in section 1.4.3) by providing additional algorithms. In Chapter 4 we applied RegrEx, $\text{RegrEx}_{\text{FVA}}$ and $\text{RegrEx}_{\text{AOS}}$ to study the central metabolism of the guard cells of the plant species *Arabidopsis thaliana*.

In contrast to the detailed discussions provided in Chapters 2 to 4, this last chapter provides a broader discussion of the results presented in this thesis. Specifically, this chapter (*i*) delineates the contributions of this thesis to the developing area of context-specific data integration in GEMs, (*ii*) discusses the challenges faced during the development of the methods presented in this thesis, and (*iii*) proposes some possibilities for improvement. The discussion will end with general considerations on context-specific data integration in GEMs, as well as possible future directions for the field.

5.1 Further considerations on Chapter 2

Chapter 2 presented a new method, termed RegrEx, to integrate transcriptomics or proteomics data into GEMs, and aimed to obtain context-specific metabolic predictions. Several methods have been proposed to integrate context-specific data into GEMs. These methods may be loosely classified into flux-centered and network-centered, based on whether they aim at obtaining context-specific flux distributions (*e.g.,* Becker & Palsson, 2008; Colijn et al., 2009; D. Lee et al., 2012; Schmidt et al., 2013) or networks (*e.g.,* Jerby et al., 2010; Schultz & Qutub, 2016; Vlassis, Pacheco, et al., 2014; Yuliang Wang et al., 2012). In this regard, RegrEx sits in between, since it was developed to provide both, context-specific flux distributions and networks.

However, two main characteristics differentiate RegrEx from other methods: fully automated and data-driven context-specific predictions. In fact, almost all existing methods require either user-defined parameters—or heuristic procedures—or the definition of some metabolic functions known to operate in a context, or both. The exception to this rule is the Lee2012 method (D. Lee et al., 2012), introduced in Chapter 2. RegrEx adds two innovations to Lee2012: Firstly, RegrEx is formulated as a single optimization problem, in which reversible reactions are directly considered by introducing binary variables. This characteristic avoids the iterative approach followed by Lee2012, which reduces the required computational time and facilitates the investigation of alternative optimal solutions (Chapter 3). Secondly, the introduction of ℓ_1-regularization allows RegrEx to obtain sparse, context-specific solutions. This is key to selecting the set of reactions that are active under a context, while eliminating, *i.e.,* shrinking the flux to zero, unspecific reactions. Altogether, these characteristics make RegrEx optimal in scenarios where little is known about the main operating metabolic functions; in these cases, a first approximation to delineating context-specific pathways must be data-driven and unguided.

The biggest challenge faced during the development of RegrEx was the formulation of the method itself. As we commented in Chapter 2, section 2.2.1, the original idea behind RegrEx was to adapt LASSO (Tibshirani, 1994), a regularized regression method, to the particular requirements of data integration in GEMs. This idea was motivated by the known benefits of ℓ_1-regularization on variable selection when dealing with high-dimensional models. A feature that could be directly applied to select important reactions under the context or interest. In addition, one could have benefited from existing efficient algorithms, which simultaneously solve the LASSO problem (2.1) for a range of λ values (Hesterberg et al., 2008). (The control parameter λ weights the effect of the regularization and must be optimized when solving the LASSO problem, see section 2.2.1). However, as shown in (2.2), the required stoichiometric and thermodynamic (*i.e.,* reaction reversibility) constraints clearly distinguish RegrEx from LASSO (2.1). This led to formulating RegrEx as the MIQP displayed in (2.4), which easily accommodates the stoichiometric constraints ($Sv = 0$) and the reversibility constraints by introducing the selecting binary variables, x.

However, this modification ruled out the possibility of using any of the existing solving algorithms to the LASSO problem, and requires solving the MIQP in (2.4) for each tested λ value. There are two directions in which RegrEx could be further improved: modifying the way in which the regularization parameter λ is selected, and including more kinds of experimental knowledge, such as metabolomics data. We will consider these two directions in more detail in the following.

In Chapter 2, we proposed the maximization of the Pearson correlation between data and flux values as our λ selection criterion (section 2.2.1, Figure 2.1). This criterion allows selecting λ in an automated and data-driven way, and provides a meaningful interpretation. However, there are other λ selection criteria that could be explored. For instance, we could balance the effect of λ on the total entropy—defined in (3.8)—of the alternative optima space of RegrEx, and the context-specific model functionality. To this end, we first need to agree on a required metabolic functionality for the context under consideration. This required metabolic functionality could be represented, for instance, as a minimum flux constraint over a selected metabolic pathway, or the production of a certain set of metabolites. We would then take samples of the alternative optima space with $\text{RegrEx}_{\text{AOS}}$ (presented in section 3.2.1.1) over a sequence of λ values, and compute the total entropy in each case. Finally, we could select the λ value that preserves the required metabolic functionality in the context-specific predictions and, at the same time, reduces the total entropy of the alternative optima space. This selection criterion implies that RegrEx predictions would no longer be entirely data-driven, since a choice of a required metabolic functionality is needed. On the other hand, it would directly address the problem of the uncertainty generated by alternative optima, and would serve to implement the balance between model sparsity and functionality discussed in section 3.2.2.

RegrEx was developed to primarily integrate transcriptomics data, since it is usually the only data source with a large reaction coverage available for most contexts. However, the integration of additional data types, when available, can render better context-specific predictions. Particularly, the consideration of several data types can further constraint the solution space of RegrEx, which reduces the alternative optima space and hence the uncertainty of the context-specific predictions. The integration of quantitative proteomics data is straightforward and can be even combined with transcript profiles during the mapping to the reactions (section 1.3.1). This approach was followed, for instance, by INIT (Agren et al., 2012), in which authors filled the "gaps" left by an incomplete protein level data set with transcriptomics data, covering as many reactions in the GEM as possible. The integration of metabolomics data in GEMs requires more elaborated strategies. However, it adds valuable context-specific information which relates more directly to metabolic reactions than transcripts or protein profiles. A simple way to integrate metabolite data is that followed by GIM^3E (Schmidt et al., 2013): first identify a set of metabolites that are known to be produced under a context, and then constrain reactions producing these metabolites to be active, *i.e.,* carry non-zero flux, in the context-specific prediction. This strategy is especially

relevant when qualitative or semi-quantitative metabolic data are available, and can be easily adopted by RegrEx—in fact, it is included in the provided RegrEx code (Robaina Estévez & Nikoloski, 2015), although it has not been evaluated in this thesis. The integration of quantitative metabolomics data requires different approaches, reviewed in Töpfer et al., (2015). RegrEx could again adopt these approaches to include both, transcript or protein levels and metabolomics data to improve the accuracy of context-specific metabolic predictions.

5.2 Further considerations on Chapter 3

The optimization problems employed in constraint-based approaches may be subject to alternative optima (see section 1.3.4). Chapter 3 explored the effects of alternative optima in the particular setting of context-specific data integration in GEMs. This thesis has made two contributions in this subject. On the one hand, the proposed methods: RegrEx$_{AOS}$ and RegrEx$_{FVA}$ (presented in Chapter 4), serve to analyze the alternative optima space of RegrEx. This analysis allows obtaining robust RegrEx predictions—as we saw in Chapter 4 when applied to the investigate the guard cell metabolism. Additionally, RegrEx$_{AOS}$ permitted the investigation of the effect of the ℓ_1-regularization on the overall uncertainty generated by the alternative optima space, as quantified by the total entropy. In turn, this investigation showed that the ℓ_1-regularization may be used to reduce the uncertainty of context-specific metabolic predictions. Moreover, while the effect of ℓ_1-regularization on variable selection is well known (Vidaurre et al., 2013), it is the first time that the added effect on reducing the alternative optima space is explored and quantified. On the other hand, the AltNet procedure, based on the context-specific network reconstruction method CorEx (presented in section 3.2.1.2), allows investigating the alternative optima space of network-centered methods aimed at reconstructing context-specific models—in this thesis we only investigated CorEx, FastCORE (Vlassis, Pacheco, et al., 2014) and CORDA (Schultz & Qutub, 2016), although it can be employed with other methods. In this case, this thesis provides the first evaluation of the effects of alternative optima on the context-specific metabolic model reconstructions obtained by this class of methods.

The biggest challenges faced during the development of the methods presented in Chapter 3 are related, here again, to the formulation of the optimization problems. In the case of RegrEx$_{AOS}$, we saw in section 3.2.1.1 and Appendix S3.1 that the original RegrEx formulation—depicted in (2.4), and renamed as RegrEx$_{OLS}$—required a modification. The alternative formulation, *i.e.*, RegrEx$_{LAD}$ (3.1), only differs in the distance function that is minimized, *i.e.,* sum of squares *vs.* sum of absolute values, and it was necessary to avoid the inclusion of a non-convex constraint (see section 1.3.4) when sampling the alternative optima space. Namely,

$$\tfrac{1}{2}\left\|\varepsilon\right\|_2^2 = Z_{opt}, \tag{5.1}$$

where $\varepsilon = d - v$ represents the error vector and $Z_{opt} = \frac{1}{2}\left\|\varepsilon_{opt}\right\|_2^2$ the sum of the squares of the optimal error vector previously found by RegrEx$_{OLS}$. Therefore, there is not available method to evaluate the alternative optima space of RegrEx$_{OLS}$. Instead, RegrEx$_{LAD}$ must be employed if we want to analyze the effects of the alternative optima on RegrEx predictions. Moreover, the differences between RegrEx$_{OLS}$ and RegrEx$_{LAD}$ predictions may be important in settings where the data vector, *d*, contains outliers (Appendix S3.1), thus it would be interesting to be able to sample the alternative optima space of RegrEx$_{OLS}$ as in the case of RegrEx$_{LAD}$. This situation may be resolved by formulating RegrEx$_{AOS}$ as a non-convex optimization problem, in which the constraint in (5.1) is included. However, non-convex optimization problems are in general harder to solve, due to the possible existence of a multitude of local optima (Boyd & Vandenberghe, 2010). A preliminary analysis, employing a state-of-the-art solver for non-convex optimization problems, has not been successful in this matter, hence this problem is still open.

On the other hand, the exploration of the alternative optima space of CorEx, and the other MBA-like methods, FastCORE, and CORDA, also brings some challenges. In this case, the biggest challenge corresponds to solving MILPs with a large number of binary variables. Binary variables are needed whenever some sort of variable exclusion constraint is required. For instance, they are employed to impose the exclusive condition on reversible reactions, where only one of the directions, either forward or backward, is allowed to carry positive flux—*e.g.,* like in the RegrEx formulation (2.4). In CorEx, apart from the set of binary variables of reversible reactions, vector *y* in (3.3), another set of binary variables, *x*, is necessary. The reason is that CorEx seeks to minimize the *support*, *i.e.,* the number of reactions carrying non-zero flux, of the non-core set *P,* as to minimize the number of non-core reactions included in the final context-specific model. Other methods, such as FastCORE, accomplish this through iterated LPs. Yet, the exploration of the alternative space, as discussed in section 3.2.1.2, requires the formulation of CorEx in a single optimization problem, which is only possible through the inclusion of binary variables. Although modern MILP solvers, such as the Gurobi solver (Gurobi Optimization, 2017), are becoming increasingly efficient, solving MILPs with a large number of binary variables is still challenging.

In this thesis, the difficulties in solving large MILPs led to abandoning a more natural formulation of the AltNet procedure (section 3.2.1.2). This formulation would follow a similar strategy to that of RegrEx$_{AOS}$ (section 3.2.1.1): first generate a random network, *i.e.,* in which the set of included non-core reactions is determined randomly, then search for the closest alternative optimal network to the previous random network, and iterate this process *n* times to obtain the sample. The optimization problem could be casted as the following QP,

$$
\begin{aligned}
&\min_{\substack{v=[v_{irr};v_{for};v_{back}]\in\mathbb{R}_0^{r+},\\ x=[x_{irr};x_{rev}]\in\{0,1\}^{|P|}\\ y\in\{0,1\}^{rev}}} \quad \tfrac{1}{2}\,\|\,\delta\,\|_2^2\\
&s.t.\\
&1-9.\ (3.3)\\
&10.\ \|\,x\,\|_1 = Z\\
&11.\ x+\delta = x_{rand},
\end{aligned}
\tag{5.2}
$$

in which: constraints 1-9 correspond to that of the optimization problem (3.3), behind CorEx, constraint 10, as in (3.4), guarantees that the alternative networks are optimal, and constraint 11 measures the distance δ between the vector of binary variables, x, indicating which non-core reactions are added, and the randomly generated binary vector, x_{rand}. However, this formulation required very large solving times, most likely due to constraint 11, which rendered it unpractical when iterated to obtain the samples.

Although the sampling procedure in (5.2) may be unpractical to solve, an alternative approach could be developed. This approach would also aim at obtaining samples of alternative optimal networks, by generating random points and finding the closest optimal network. However, instead of generating the random point as a binary vector, x_{rand}, this procedure would create a random flux vector, v_{rand}, and then search for the closest point in the flux cone K (1.21) which renders the optimal support in the set of non-core reactions. Again, this optimization problem could be casted as the QP,

$$
\begin{aligned}
&\min_{\substack{v=[v_{irr};v_{for};v_{back}]\in\mathbb{R}_0^{r+},\\ \delta=[\delta_{irr};\delta_{rev}]\in\mathbb{R}^{r},\\ x=[x_{irr};x_{rev}]\in\{0,1\}^{|P|}\\ y\in\{0,1\}^{rev}}} \quad \tfrac{1}{2}\,\|\,\delta\,\|_2^2\\
&s.t.\\
&1-9.\ (3.3)\\
&10.\ \|\,x\,\|_1 = Z\\
&\left.\begin{aligned}
&11.\ v_{irr(i)} + \delta_{irr} = v_{rand}\\
&12.\ (v_{for(i)} - v_{back(i)}) + \delta_{rev} = v_{rand(i)}
\end{aligned}\right\},\quad i\in P.
\end{aligned}
\tag{5.3}
$$

In (5.3), constraints 1-9 are again inherited from the optimization problem (3.3), and constraint 10 guarantees the optimality of the alternative networks. Constraints 11 and 12 measure the distance, δ, to v_{rand}, the randomly generated vector of flux values for the non-core reactions (*i.e.,* the indexes $i\in P$ in the GEM). Constraint 11 accounts for irreversible reactions while constraint 12 for reversible reactions.

Besides the developing of new methods, future improvements should also be tailored towards elaborating decision procedures to select a representative of the alternative space. In section 3.2.2.4, we proposed a possible selection procedure, in which

reactions were first ranked according to their frequency in the alternative optima space, and then preferentially included in the final context-specific model according to this ranking. Moreover, we discussed the possibility of using (post-optimization) metabolic tests, in which the space of alternative models is reduced by selecting those that outperform at certain predefined metabolic tasks. Both strategies represent good starting points to solve the problem of uncertain context-specific metabolic predictions due to alternative optima.

5.3 Further considerations on Chapter 4

Chapter 4 presented a direct application of $RegrEx_{LAD}$ and $RegrEx_{AOS}$ to investigate the central metabolism of the guard cells of *Arabidopsis thaliana.* The major contributions of this study were threefold: (*i*) provide large-scale metabolic predictions specific to guard cell and mesophyll cells, (*ii*) a robust differential analysis between the two cell types, in which the alternative optima space of $RegrEx_{LAD}$ was considered, and (*iii*) an independent ^{13}C experiment conducted to validate model predictions. These contributions differentiate this study from others which focused on modeling the dynamical (Hills et al., 2012; Minguet-Parramona et al., 2016) and the signaling (Li et al., 2006; Sun et al., 2014) processes controlling stomatal opening, and which did not analyze the metabolic differences between the two cell types.

The major challenges encountered during the development of this study involved the lack of biological knowledge on the guard cell physiology, as well as its relations with the mesophyll cells. On the one hand, this scenario aligns with the purpose of RegrEx, since fully data-driven predictions are the only choice when little knowledge is available for the context of interest. On the other hand, the data-driven predictions provided in this study require further experimental evaluation, and may be regarded as a first approximation to delineating the context-specific metabolic activity in the guard cells. For instance, future studies could explore the effects of including transport reactions, such as malate exchange, between mesophyll and guard cells. This inclusion could shed light on the role of mesophyll cells as possible carbon sources to guard cells (Araújo et al., 2011; Nunes-Nesi et al., 2007; Penfield et al., 2012), as well as their relative contributions to the total carbon pool. However, this addition would require the experimental determination of, at least, upper flux bounds on these transport reactions. Otherwise, the conclusions obtained by the modeling predictions could be biased.

Besides better and more complete experimental observations, the *in silico* investigation of guard cell physiology would benefit from a combined modeling strategy. In this setting, GEMs could be combined with dynamical models of stomatal function: GEMs would serve as a framework to integrate high-throughput data and to obtain flux predictions, in contrast, the dynamical models would capture the signaling and electrophysiological processes controlling stomatal opening. The two modeling frameworks could be connected through key reactions, for instance, the synthesis of

malate or sucrose, which are included in the OnGuard dynamical model (Hills et al., 2012). Furthermore, a similar approach as that of Dynamic Flux Balance Analysis (Radhakrishnan Mahadevan, Edwards, & Doyle, 2002) could be followed to accommodate dynamic predictions in the static modeling framework of the GEM. Namely, discretize time into time intervals, assume quasi-steady state of the metabolic network within the intervals, and use the dynamic models to constraint the flux through key input reactions of the GEM at the beginning of each time interval. As commented in section 4.1, this strategy has been successfully employed to investigate the evolution of C_4 photosynthesis[5] (Mallmann et al., 2014), and could be readily adapted to the guard cell scenario.

5.4 Future directions

Context-specific metabolic model predictions would benefit from further developments in two areas. On the one hand, improvements in the accuracy of GEMs, the quality of experimental data, and further developments of the computational methods[6] used to combined these elements would naturally lead to better predictions. On the other hand, new elements could be added to the basic schema of GEMs plus high-throughput data followed by current methods—such as the consideration of non-metabolic and regulatory cellular processes. In the following, we will discuss some of the key points in both areas.

Any improvement in the reconstruction of GEMs leading to better metabolic representations would benefit context-specific predictions—all constraint-based predictions for that matter. However, there are two elements related to the experimental data mapping to reactions (section 1.42) that are key to context-specific predictions: gene-protein-reaction rules and gene, reaction and metabolite canonical identifiers. Gene-protein-reaction rules are crucial to integrating transcriptomics and proteomics data into GEMs. Not only because they are required to map these data into the reactions, but also because the experimental values mapped to the reactions depend on the accuracy of the rules and on the assumptions taken during their elaboration. Yet, not all available GEMs are equipped with gene-protein-reaction rules. In other occasions, the rules suffer from bias due to the automated generation by algorithms that do not consider key biological constraints, such as the compartment-specific representation of certain enzymes or transporters. Although there is active research in this issue (Machado, Herrgård, & Rocha, 2016), further improvements in the representation of gene-protein-reaction rules would markedly impact the correctness of context-specific predictions.

[5] In this case, a dynamical model describing the transport processes between mesophyll and bundle sheath cells was combined with a GEM of a C_4 plant, which included both cell types.

[6] The consideration of alternative optima during context-specific metabolic predictions constitutes an immediate improvement; we will not cover it here since we discussed about this in the previous section.

Furthermore, the integration of both, transcriptomics and proteomics data, and metabolite profiles requires unambiguous canonical identifiers for these elements. This is particularly important for automated mappings between experimental values and model elements—*i.e.,* reactions and metabolites—which are free from human errors and save huge amounts of time. GEMs usually provide canonical gene identifiers, this is not always the case with protein names, and it is even less frequent with metabolite names, situation that greatly impairs the integration of metabolomics data. While the improvement of gene-protein-reaction rules requires considerable research effort, including canonical identifiers is only a matter of convention, hence readily solvable.

Besides enhancing GEMs and improving the computational methods, the inclusion of additional, non-metabolic elements would expand the accuracy and biological relevance of context-specific metabolic predictions. In this area, we have already commented the possibility of coupling small-scale dynamical models to GEMs, as to obtain dynamical and large-scale predictions (section 5.3). Another route of expansion consists of considering regulatory elements during context-specific predictions. In this sense, the inclusion of gene regulatory networks[7] into constraint-based metabolic modeling is relatively well explored (Covert, 2002; Covert, Schilling, & Palsson, 2001; Faria et al., 2013; S. Ma et al., 2015; Marmiesse, Peyraud, & Cottret, 2015) and can be readily employed to enhance context-specific metabolic predictions. Specifically, gene regulatory networks constrain the flux through the reactions mapped to regulated genes in response to input effectors, such as the presence or absence of a metabolite or a transcription factor. These regulatory processes are dynamical, and hence require dynamic simulations. In this case, as when integrating small-scale dynamic models, dynamic simulations can be realized by subdividing the time coordinate into time-steps, and assuming quasi-steady-state within each time-step. GEM predictions are then updated at the beginning of each time-step by including the specific constraints on reaction flux bounds, as processed by the gene regulatory network. We could then render the metabolic predictions context-specific by simultaneously considering context-specific gene regulatory networks, *i.e.,* specifically developed for the context of interest (Geeven, van Kesteren, Smit, & de Gunst, 2012; Koch, 2016), and further constraining GEM predictions by integrating context-specific data.

Finally, a complete understanding of the physiological processes governing life requires the development of multi-scale models. In this scenario, diverse

[7] Gene regulatory networks represent all known transcriptional interactions between the genes of a species. Typically, these interactions take place between regulatory elements, the transcription factors, and their target genes. The transcription factors, proteins coded by regulatory genes, can repress of activate the transcription rate, *i.e.,* the expression, of their target genes. At the same time, the effect and even the nature of the interaction depend on particular inputs that are affect the transcription factors. The reconstruction of gene regulatory networks from time series data of transcript levels is an active area of research, and a variety of computational methods to perform this task is available (Chai et al., 2014)

physiological processes occurring at various time-scales and locations in an organism are modeled. A network of cell-specific metabolic models can serve as a base framework for the multiscale model. These context-specific models can then be coupled through transport reactions that, in turn, may be described by small-scale dynamical models. On top of this basic skeleton, gene regulatory networks coupled with dynamic models of key signaling processes can be used to constraint the flux outputs of the metabolic models. Naturally, multi-scale models of this sort are ambitious and difficult to obtain. However, the first steps toward this direction have already been taken (Gomes de Oliveira Dal'Molin, Quek, Saa, & Nielsen, 2015; Grafahrend-Belau et al., 2013; Karr et al., 2012), and current methods and technologies grant ample room for improvement.

Supplementary Information

Chapter 2

The supplementary information of Chapter 2 can be retrieved from:

Robaina-Estévez, S., Nikoloski, Z. *Context-specific metabolic model extraction based on regularized least squares optimization.* PLOS ONE (2015). 10.1371/journal.pone.0131875

Chapter 3

The supplementary information of Chapter 3 can be retrieved from:

Robaina-Estévez S, Nikoloski Z. On the effects of alternative optima in context-specific metabolic model predictions. PLoS Comput Biol. 2017;13. doi:10.1371/journal.pcbi.1005568

Chapter 4

The supplementary information of Chapter 4 can be retrieved from:

Robaina Estévez S, Daloso DM, Zhang Y, Fernie AR, Nikoloski Z. Resolving the central metabolism of Arabidopsis guard cells. Sci Rep. 2017, 10.1038/s41598-017-07132-9

Appendix S3.1. Detailed description of RegrEx$_{LAD}$ and comparison with the original RegrEx$_{OLS}$

Here we justify the usage of RegrEx$_{LAD}$, presented in the MILP (3.2) instead of the original RegrEx version (renamed as RegrEx$_{OLS}$), in MIQP (2.4), when analyzing the alternative optima space. As commented in Chapters 2 and 3, the RegrEx$_{OLS}$ method minimizes the squared ℓ_2 norm of $\epsilon = d - v$, the difference vector between the experimental data vector, d, and a feasible flux distribution, v. This is indeed the only difference with respect to the RegrEx$_{LAD}$ method, which minimizes the ℓ_1 norm (*i.e.*, the sum of absolute error values) of ϵ.

During the course of this study, we first tried to investigate the alternative optima space of RegrEx$_{OLS}$ through a sampling procedure, in a way akin to the Variability Flux Sampling procedure implemented in (Recht et al., 2014). The Variability Flux Sampling procedure was developed to investigate the alternative optima space of the InGenMinimizer method (presented in the same publication), by generating a random sample of alternative optimal flux distributions. The InGenMinimizer method follows the quadratic program,

$$\begin{aligned} & Z_{opt} = \min_{v,\varepsilon} \tfrac{1}{2} \| \varepsilon \|_2^2 \\ & s.t. \\ & 1.\ Sv = 0 \\ & 2.\ v_{\min} \le v \le v_{\max} \\ & 3.\ v_i = d_i + \varepsilon_i, \forall i \in R_D. \end{aligned} \qquad \text{(A.1)}$$

Therefore, RegrEx$_{OLS}$ can be seen as an extension of the InGenMinimizer method (A.1), in which (*i*) ℓ1-regularization is included in the objective function, and (*ii*) reversible reactions with associated data are also considered in the minimization, which requires introducing a vector of binary variables, x (as explained in the main text).

The Variability Flux Sampling procedure was formulated as the quadratic program,

$$\begin{aligned} & \min_{v,\varepsilon,\delta} \tfrac{1}{2} \| \delta \|_2^2 \\ & s.t. \\ & 1.\ Sv = 0 \\ & 2.\ v_{\min} \le v \le v_{\max} \\ & \left.\begin{aligned} & 3.\ v_i = d_i + \varepsilon_i \\ & 4.\ \tfrac{1}{2} \| \varepsilon \|_2^2 = Z_{opt} \end{aligned}\right\}, \forall i \in R_D \\ & 5.\ v = v_{rand} + \delta \end{aligned} \qquad \text{(A.2)}$$

which minimizes the distance, δ, between an alternative optimal flux distribution, v, and a randomly generated flux distribution, v_{rand}. The QP in (A.2) is solved n times to obtain a sample of n alternative optimal flux distributions). The key in (A.2) is the constraint number 4, *i.e.*, $\frac{1}{2}\|\epsilon\|_2^2 = Z_{opt}$,which guarantees that any sampled feasible v, is also optimal, since it renders the same squared ℓ_2 norm of ϵ previously obtained by (A.1). This is a quadratic equality constraint, which makes (A.2) non-convex and thus intractable by convex optimization tools. Note that this constrain would also be required in the case of RegrEx$_{OLS}$, since it also minimizes the squared ℓ_2 norm of ϵ.

In the Variability Flux Sampling procedure, authors used a non-convex solver, MINOS (Bruce A. Murtagh, n.d.), to tackle this problem. However, several aspects make the case of RegrEx$_{OLS}$ more complex: firstly, in the Variability Flux Sampling procedure authors only dealt with seven reactions with associated data, in contrast, RegrEx$_{OLS}$ must evaluate all reactions with associated data in a GEM. Secondly, integer variables were not required in the Variability Flux Sampling procedure, since all seven reactions were irreversible, as oppose to RegrEx$_{OLS}$, where reversible reactions with associated data are also consideredt. Lastly, a flux distribution that is alternatively optimal to RegrEx$_{OLS}$ must also render the same ℓ_1 norm as the original optimum, thus a second constraint like $\|v_s\|_1 = \|v_{opt}\|_1$ must be added. Altogether, these particularities make the optimization problem associated to any RegrEx$_{OLS}$ alternative optima sampling procedure hardy tractable by any existing solver. However, it is computationally tractable to sample alternative optimal solutions of RegrEx$_{LAD}$. This is because the objective function of RegrEx$_{LAD}$ is linear, and hence only two linear constraints are required to guarantee that a sampled flux distribution is optimal to RegrEx$_{LAD}$. Thus, the sampling procedure (RegrEx$_{AOS}$, see main text) can be casted as a convex optimization problem and solved with existing solvers.

Although computational tractability was our main motivation to develop RegrEx$_{LAD}$, we noted that this alternative version may have another advantage over the original RegrEx$_{OLS}$. RegrEx$_{OLS}$ and RegrEx$_{LAD}$ parallel two classical approaches followed in linear regression, namely, the ordinary least squares (OLS) and the least absolute deviations (LAD, also known as least absolute value, LAV) method (Lawrence & Shier, 2010). OLS and LAD regression behave differently upon the presence of outliers in the distribution of errors (*i.e.*, the vector ϵ), that is, elements that are very far away from the mean of the distribution. Concretely, the OLS method tends to get biased results in such cases, since the squared ℓ_2 norm of ϵ gives excessive importance to these elements. On the other hand, the LAD method is more robust under the presence of such outliers, and thus less prone to give biased results (in fact, the LAD method is the simplest among the so-called Robust Regression techniques, see for instance (Dielman, 2005)). In the context of RegrEx, this means that RegrEx$_{LAD}$ could be a better choice in cases where outliers are present in the error distribution, for instance, if a given mapped gene expression value is particularly high with respect to the mean value of the gene expression data set. In fact, this idea has been implemented in the case of the least absolute shrinkage and selection (LASSO) operator (Tibshirani, 1994) (which inspired

the development of RegrEx$_{OLS}$), which applies a ℓ_1-regularization to an OLS regression. Concretely, the LASSO has been adapted to a LAD regression, showing advantages in cases where the distribution of errors is not appropriate for OLS estimation (H. Wang, Li, & Jiang, 2007).

To test the previous idea, we evaluated the RegrEx$_{OLS}$ and RegrEx$_{LAD}$ performance under the inclusion of outliers in the leaf data set used in the main study. To this end, we first generated a sample of randomly perturbed leaf data vectors, $d^*_{Leaf(j)} = d_{Leaf} + \mu_{(j)},\ j = \{1, \dots, 10^4\}$, obtained by adding a uniform noise ,$\mu_{(j)}$, (±1% of the mean value of d_{Leaf}) to the original leaf data set, d_{Leaf}. We next obtained a "contaminated" leaf data set, which contained an outlying expression value for one of the reactions. Concretely, we substituted the data associated to the reaction that had the minimum value in d_{Leaf} by a large amount, in this case 5 times the maximum value in d_{Leaf}. We then applied RegrEx$_{OLS}$ and RegrEx$_{LAD}$ using the AraCORErd model and the contaminated leaf data set, and calculated the total sum of the absolute errors,

$$T_{\varepsilon(j)} = \frac{1}{|R_D|}\sum_j v_{(i,j)} - d^*_{Leaf\,(i,j)}, \qquad \text{(A.3)}$$

with $i = \{1, \dots, R_D\}$, where $|R_D|$ corresponds to the number of reactions with associated data, between the optimum RegrEx$_{OLS}$ and RegrEx$_{LAD}$ flux distributions and each of the perturbed leaf datasets, $d^*_{Leaf(j)}$, in the sample (the code used in this evaluation is included in File S3.1). In this evaluation, RegrEx$_{LAD}$ rendered smaller total sums of absolute errors across the perturbed data sample (mean $T_\epsilon = 0.718$ in RegrEx$_{LAD}$ *versus* a mean $T_\epsilon = 0.722$ in RegrEx$_{OLS}$) as determined by a two-sided Mann-Whitney test (p-value = 0). In addition, RegrEx$_{LAD}$ did not render smaller total sums when the original ("uncontaminated") leaf data set was used under the same setting (mean $T_\epsilon = 0.709$ in RegrEx$_{LAD}$ *versus* a mean $T_\epsilon = 0.706$ in RegrEx$_{OLS}$, p-value = 1) . Although the reported differences between total sums of absolute errors are small, they serve to illustrate the more robust behavior of RegrEx$_{LAD}$ under the presence of outliers.

Appendix S3.2. Description of iMAT$_{AOS}$ and application to the two evaluated case studies

Alternative optimal solutions in the iMAT method: background

As mentioned in the main text, there already exists a procedure to investigate the alternative optima space for iMAT (Shlomi et al., 2008). Hence, we considered relevant to apply iMAT to the same context-specific reconstructions examples used in CorEx, and analyze its alternative optimal solutions. Here, we briefly summarize the iMAT method as well as the procedure proposed by the authors to analyze its alternative optima. In addition, we present our novel complementary approach to sample the alternative optima space of iMAT. iMAT aims at maximizing the global similarity between a given expression data set and a feasible flux distribution of the GEM where data is being integrated. Therefore, in this sense, it follows an approach similar to RegrEx. However, iMAT does not directly minimize the distance between data and flux values. Instead, iMAT first integrates experimental information by classifying reactions in the GEM into two groups: one, R_H, populated by reactions with *highly* expressed associated genes (*i.e.*, above a fixed threshold value, ϵ) and another, R_L, by reactions with *lowly* expressed associated genes (*i.e.,* below ϵ). The MILP presented in (A.4) is then solved to maximize the number of active reactions in R_H (with non-zero flux) and the number of inactive reactions in R_L (with zero flux value), subject to the usual mass balance and thermodynamic constraints. This is implemented by maximizing the norm of two vectors of binary variables, y^+, y^-, that select reactions in R_H to be active and reactions in R_L to be inactive (the extra variable, y^-, is added to account for reversible reactions).

$$
\begin{aligned}
&Z_{opt} = \max_{\substack{v\in\mathbb{R}^m \\ y^+,y^-\in\{0,1\}}} \sum_{i\in R_H}(y_i^+ + y_i^-) + \sum_{i\in R_L} y_i^+ \\
&s.t. \\
&1.\ Sv = 0 \\
&2.\ v_{\min} \le v \le v_{\max} \\
&3.\ v_i + y_i^+(v_{\min,i} - \varepsilon) \ge v_{\min,i}, i \in R_H \\
&4.\ v_i + y_i^-(v_{\max,i} + \varepsilon) \le v_{\max,i}, i \in R_H \\
&5.\ v_{\min,i}(1 - y_i^+) \le v_i \le v_{\max,i}(1 - y_i^+), i \in R_L
\end{aligned}
\qquad \text{(A.4)}
$$

To deal with alternative optimal flux distributions, authors (Shlomi et al., 2008) proposed the following approach, which we denominate here iMAT$_{FVA}$. First, the MILP in (A.4) is solved twice for each reaction in the GEM; the first time, the reaction is forced to be active, in the second, to be inactive. The two objective values, $Z_{act(i)}$, $Z_{inac(i)}$, corresponding to the optimizations where reaction i was active and inactive, respectively, are then compared. If $Z_{act(i)} > Z_{inac(i)}$, reaction i is considered to be active (with confidence $Z_{act(i)} - Z_{inac(i)}$), if $Z_{act(i)} < Z_{inac(i)}$, is considered to be inactive (with confidence $Z_{act(i)} - Z_{inac(i)}$) and if $Z_{act(i)} = Z_{inac(i)}$ is taken as undetermined under the data

set been integrated. Therefore, iMAT$_{FVA}$ determines the sets of reactions that individually increase the global similarity to data when active and inactive, respectively, and the set of reactions that do not affect the optimum global similarity to data under whatever state, active or inactive. However, it does not provide information about how the states of reactions distribute across the alternative optima space of iMAT. For instance, a given reaction could be classified as active and still be either active or inactive across the space of all alternative optimal flux distributions generated by iMAT.

We emphasize that the results obtained through iMAT$_{FVA}$ do not align qualitatively to the ones obtained for RegrEx and CorEx (by extension FastCORE and CORDA), and hence they have to be interpreted on their own. To make a fair comparison, we need a method that allows drawing samples of alternative optimal flux distributions to iMAT. In the EXAMO publication (Rossell et al., 2013), authors generated such a sample by collecting the flux distributions that rendered the maximum objective value, Z_{opt} in (A.5), when applying iMAT$_{FVA}$. Therefore, we can only generate a limited number of sampled optimal flux distributions with this method. Here, we propose a different procedure to evaluate the alternative optimal space of iMAT (iMAT$_{AOS}$), which follows a similar approach to the one employed by RegrEx$_{AOS}$: we first generate a random flux distribution, v_{rand}, and then search for the closest feasible flux distribution, v, that renders the same optimal result, Z_{opt}, found by iMAT. iMAT$_{AOS}$ optimizes the mixed integer quadratic program:

$$\min_{\substack{v,\delta\in\mathbb{R}^m \\ y^+,y^-\in\{0,1\}}} \ \tfrac{1}{2}\|\delta\|_2^2$$

$$s.t.$$

$$1-5 \quad \text{(A.4)} \qquad \text{(A.5)}$$

$$6.\ v = v_{rand} + \delta$$

$$7.\ \sum_{i\in R_H}(y_i^+ + y_i^-) + \sum_{i\in R_L} y_i^+ = Z_{opt}$$

The MIQP in (A.5) inherits constraints 1-5 from (A.4) and includes constraint 6, which defines the distance, $\delta = v - v_{rand}$ to be minimized, and constraint 7, which guarantees that v remains within the alternative optimal space of the previous iMAT optimization. In this manner, iMAT$_{AOS}$ allows drawing an unlimited number of random alternative flux distributions that are optimal to (A.4).

Alternative optimal solutions in the iMAT method: case studies

We next applied iMAT and iMAT$_{FVA}$—to analyze its alternative optimal solutions—to AraCORErred and Recon1red. In this case, we used the core set of reactions for the leaf and the liver contexts as the R_H group in iMAT. In this way, we obtained a leaf-specific model containing 131 reactions and 154 metabolites, while the liver-specific model consisted of 1235 reactions and 1067 metabolites. By applying iMAT$_{FVA}$, we found a total of 272 active, 178 inactive and 5 undetermined reactions across the iMAT

alternative optima space for the leaf context. For the liver context, the alternative optima space included 1223 active, 981 inactive and 143 undetermined reactions in the case of liver (Table A.1). We quantified the uncertainty of the iMAT data integration problem by taking the proportion of undetermined reactions over the total number in the GEM. The undetermined reactions in the alternative optima space for the leaf and the liver-contexts were 1.1% and 6.1%, respectively.

	LEAF				LIVER			
	A	**I**	**U**	$\overline{M_R}$ **(CV)**	**A**	**I**	**U**	$\overline{M_R}$ **(CV)**
IMAT$_{FVA}$	272	178	5	-	1223	981	143	-
IMAT$_{AOS}$	275	40	140	43.16(0.20)	1069	247	1153	369.32(0.05)
OVERLAP	259 (95.2%)	35 (19.7%)	0	-	928 (75.9%)	69 (5.6%)	2 (0.16%)	-

Table A.1. Summary of the alternative optima space of iMAT. This table includes the number of active, A, inactive, I, and undetermined, U, reactions across the alternative optima space as determined by iMAT$_{FVA}$ and the iMAT$_{AOS}$. The intersection between the two methods is also displayed for each of the three categories (Overlap). Finally, the mean number of reaction mismatches (*i.e.*, the Hamming distance), $\overline{M_R}$, between the generated alternative optimal networks (see main text) is also displayed (the coefficient of variation, CV, is shown in parenthesis). These figures are displayed for the leaf- and the liver-specific scenario.

We next evaluated the alternative optimal space with iMAT$_{AOS}$, which allowed us to draw two random samples (size n = 2000) of leaf- and liver-specific alternative optimal flux distributions. We focused on characterizing the state of the reactions, as active or inactive, across the sample. For the leaf context, 60% of the reactions were active in all alternative flux distributions, 9.23% had a fixed inactive state, and a 30.8% were of undetermined state across the sample. For the liver context, the fraction of fixed active reactions amounted to a 43.3%, 9.52% showed fixed inactive state, and 47.2% of the reactions were of undetermined state across the sample (Table A.1). Here, too, we considered the fraction of reactions with undetermined state across the alternative optima sample as an uncertainty measure of the iMAT data integration problem. Our results demonstrated that the uncertainty for the liver context was greater than that for the leaf (Table A.1), which agrees with the results previously obtained in the case of CorEx. These findings were supported by the significantly different Hamming distance calculated between any possible pair of alternative optimal networks (one-sided ranksum test, p-value = 0, see Methods, section 3.4).

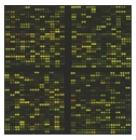
Additionally, a comparison of the results obtained through the two alternative methods, iMAT$_{FVA}$ and iMAT$_{AOS}$, showed a good agreement in the sets of reactions classified as active across the alternative optima space: a 94.8% of active reactions per iMAT$_{FVA}$

were also found active by $iMAT_{AOS}$ in the leaf context, and a 75.9 % for the liver context. However, this agreement did not hold in the case of inactive and undetermined reactions, both in leaf and in liver (Table A.1). Therefore, this comparison highlighted the importance of analyzing a sample of alternative optimal solutions to obtain a more complete understanding of the uncertainty associated to an experimental data integration problem.

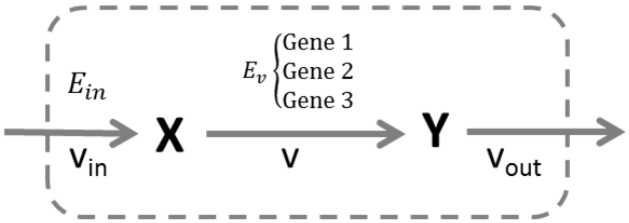
iMAT implementation and alternative optima evaluation

The iMAT implementation was taken from the function *createTissueSpecificModel* in the COBRA toolbox (D. Hyduke et al., 2011) (for MATLAB) and slightly modified to allow the usage of the Gurobi solver (version 7.01), used throughout this study. In addition, the $iMAT_{FVA}$ procedure was performed through adapting the previous iMAT implementation (no publicly available implementation of this procedure was found). Both MATLAB functions can be found in File S3.1 under the names of *iMAT* and $iMAT_{FVA}$. In addition, the implementation of our alternative sampling method, $iMAT_{AOS}$, can be found in the same file.

1) Preprocess data



gene name – RMA value
Gene 1– 1.1
Gene 2– 1.4
Gene 3 – 1.5



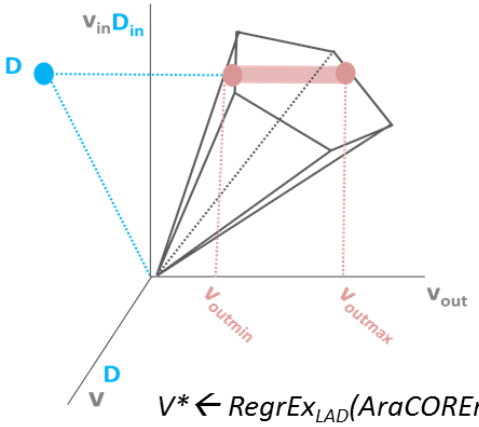
RMA (affy R package)
Affymetrix probe names → Arabidopsis gene names

2) Map data to reactions

***Ev rule*:** Gene 3 OR Gene 1 OR (Gene 3 AND Gene 2) OR (Gene 1 AND Gene 2)
Dv: max(Gene 3, Gene 1, min(Gene 3, Gene 2), min(Gene 1, Gene 2)) = 1.5

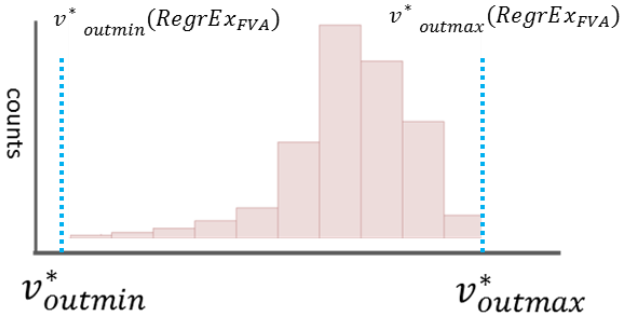
D ← mapgene2rxn(GPR rules, gene names, RMA value)

3) Obtain cell-specific flux distribution



V ← $RegrEx_{LAD}$(AraCORERed,D)*

4) Obtain AO space sample



V_{AO} ← $RegrEx_{AOS}$(AraCORERed,D,V)*

5) Obtain flux distributions per metabolite and compare between cell-types

h_1 (MWW test, α = 0.05):

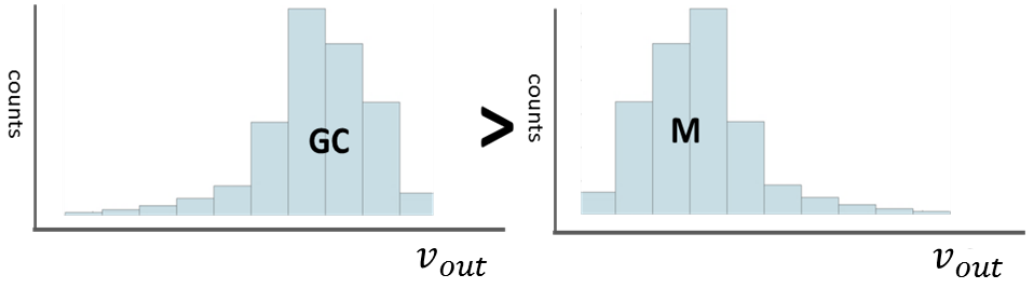


Figure S4.1. Schematic depiction of the workflow followed to obtain metabolic predictions specific to G and M cells. *This schematic depiction is based on the toy metabolic model displayed in the top right: X first enters the system through the reaction, v_{in}, which is dependent on the transporter E_{in}. X is then transformed to Y through v, which is dependent on E_v (coded by gene1-3). Finally, Y diffuses spontaneously to the exterior (v_{out}), and hence no genes are associated to this reaction. (1) In a first step, expression data is preprocessed, which includes data RMA normalization and mapping of array probe names to (Arabidopsis) gene names. (2) Gene expression values are then mapped to reactions in the metabolic model following the gene-protein-reaction rules contained in the model, which generates the vector D of mapped expression values. (3) In a third step, a cell-specific flux distribution that is closest to the mapped expression data integrated in the metabolic model is obtained through $RegrEx_{LAD}$. However, the optimal solution (i.e.,flux distribution) is not unique, and an alternative optima space (AO) exists. In this case, this is because D contains data to only two of the three reactions (v_{in} and v) since v_{out} has not associated gene-protein-reaction rule. Thus, v_{out} can vary in an orthogonal direction to the plane, where D lies, without affecting the optimal value of the $RegrEx_{LAD}$ objective function (section 3.2). (4) To account for this issue, the AO space is sampled through $RegrEx_{AOS}$ and a sampled distribution of optimum flux values for v_{out} is obtained. Additionally, the function $RegrEx_{FVA}$ calculates the minimum and maximum values in the alternative optima space, as a means to validate the coverage of the random sample. (5) Finally, the resulting AO flux and flux-sum distributions are compared through a Mann-Whitney-Wilcoxon (MWW) test. In this example, the alternative hypothesis (H_1) states that the distribution of flux values corresponding to G cells is greater than that of M cells, while the null hypothesis states that both distributions are indistinguishable. The null hypothesis is rejected at a significant level of $\alpha = 0.05$.*

Table S4.1. A comparison of the predicted metabolic state of GC and M. The predicted mean flux values corresponding to the reactions depicted in Figure 4.1 are shown for G and M cells. The p-values shown in this table correspond to a Mann-Whitney test comparing the distributions of flux values of G and M cells (V_G and V_M, respectively). Three different tests were considered for each comparison: (i) we evaluated whether the two distributions differ (null hypothesis Ho: $V_G = V_M$, (ii) if G had increased flux values in comparison to M cells (Ho: $V_M > V_G$), and (iii) if M had increased flux values in comparison to G cells (Ho: $V_G > V_M$). In all three cases, a horizontal bar indicates a failure of the test due to distributions consisting of a fixed value. Reaction names and index numbers in accord with AraCOREred.

Idx. in Fig.1	Reaction Name (Idx in AraCOREred)	Mean Flux G	Mean Flux M	Mean ratio (G/M)	Ho: $V_G = V_M$	Ho: $V_M > V_G$	Ho: $V_G > V_M$	SubSystem
PSII	photosystem II (1)	0.125	0.125	1.000	0.317	0.159	0.841	light reactions
Cb6f	cytochrom b6f complex (2)	0.250	0.250	1.000	0.317	0.159	0.841	light reactions
PSI	photosystem I (3)	0.500	0.500	1.000	0.317	0.159	0.841	light reactions
ATPase	ATPase (5)	0.107	0.107	1.000	0.317	0.159	0.841	light reactions
1	PEP carboxylase (54)	0.420	0.034	12.486	0	0	1.000	gluconeogenesis
2	Cytosolic NADP-MDH (115)	0.161	0.084	1.917	0	0	1.000	pyruvate metabolism
3	Dicarboxylate transporter (339)	0.251	0.033	7.554	0	0	1.000	transport
4	Plastidial NADP-Malic Enzyme (113)	0.249	0.032	7.848	0	0	1.000	pyruvate metabolism
5	CO2 diffusion [Forward] (374)	0	0.001	0	0	1.000	0	transport
5	CO2 diffusion [Backward] (374)	0.230	0.013	17.342	0	0	1.000	transport
6	Import CO2 (413)	0.021	0.021	1.000	0.013	0.994	0.006	import
7	Carbonic anhydrase (152)	0.420	0.034	12.486	0	0	1.000	carbon fixation
8	Glu tamate synthase (FeS-Fd) (179)	0.124	3.671e-05	3.391e+03	0	0	1.000	glutamate synthesis
9	Ferredoxin-NADP reductase (4)	0.126	0.250	0.502	0	1.000	0	light reactions
10	Export O2 (420)	0.020	0.020	1.000	0.847	0.424	0.576	export
11	TP isomerase [Forward] (9)	0.854	0.855	0.999	0	1.000	0	Calvin-Benson cycle, glycolysis
12	TP/Pi translocator [Forward] (328)	0.857	1.000	0.858	0	1.000	0	transport
13	Di-/ri-carboxylate carrier [Forward] (346)	0.348	0.327	1.065	0	0	1.000	transport
13	Di-/ri-carboxylate carrier [Forward] (347)	0.345	0.324	1.065	0	0	1.000	transport
13	Di-/ri-carboxylate carrier [Forward] (348)	0.349	0.324	1.077	0	0	1.000	transport
14	Di-/ri-carboxylate carrier [Backward] (343)	0.311	0.294	1.059	0	0	1.000	transport
14	Di-/ri-carboxylate carrier [Backward] (344)	0.314	0.289	1.088	0	0	1.000	transport
14	Di-/ri-carboxylate carrier [Backward] (345)	0.309	0.292	1.057	2.395e-06	1.198e-06	1.000	transport
15	Mitochondrlial NAD-MDH [Backward] (80)	0.271	0.226	1.201	0	0	1.000	tricarboxylic acid cycle, glyoxylate cycle
15	Mitochondrial NADP-MDH (117)	0.753	0.724	1.041	0	0	1.000	pyruvate metabolism
16	FBP aldolase [Forward] (35)	0.018	0.160	0.112	0	1.000	0	sucrose synthesis, gluconeogenesis, glycolysis
17	FBPase (36)	0.073	0	Inf	0	0	1.000	sucrose synthesis, gluconeogenesis, glycolysis
17	PPi-dep. Phosphofructokinase [Backward] (136)	0	0.693	0	0	1.000	0	pyrophosphate recycling
18	G6P isomerase [Forward] (39)	0.018	0.160	0.112	0	1.000	0	sucrose synthesis, sucrose degradation, gluconeogenesis
19	Phosphoglucomutase [Forward] (40)	0.055	7.590e-04	72.503	0	0	1.000	sucrose synthesis, sucrose degradation
20	TP/Pi translocator [Forward] (327)	0	0	0	-	-	-	transport
20	TP/Pi translocator [Backward] (327)	0.362	0.439	0.825	0	1.000	0	transport

Table S4.2. Predicted Flux-Sums of selected metabolites in the AraCOREred model. The predicted mean flux-sum values for several metabolites are displayed. In each of the metabolites, the flux-sum values are split into each cellular compartment: cytosol (c), mitochondrion (m), peroxisome (p) and chloroplast (h). In addition, the total flux-sum values (*i.e.,* taking into account all the compartments) are provided. The interpretation of the p-values is similar to that of Table S1. A horizontal bar indicates a failure of the test due to distributions consisting of a fixed value.

Metabolite (compartment)	Mean FluxSum G	Mean FluxSum M	Ratio (G/M)	Ho: $V_G = V_M$	Ho: $V_M > V_G$	Ho: $V_G > V_M$
Total Mal	5.434	4.581	1.186	0	0	1.000
Mal(c)	2.451	2.172	1.128	0	0	1.000
Mal(m)	2.129	2.004	1.062	0	0	1.000
Mal(p)	0	0	-	-	-	1.000
Mal(h)	2.211	1.493	1.481	0	0	1.000
Total Suc	0.109	5.574e-05	1.948e+03	0	0	1.000
Futile Cycle Suc	0.109	1.203e-05	9.019e+03	0	1.000	0
Total OAA	7.460	6.729	1.109	0	0	1.000
OAA(c)	4.063	3.606	1.127	0	0	1.000
OAA(m)	2.083	1.950	1.069	0	0	1.000
OAA(p)	1.320	1.488	0.887	0	1.000	0
OAA(h)	1.713	1.430	1.198	0	0	1.000
Total Pyr	3.571	3.085	1.158	0	0	1.000
Pyr(c)	1.418	1.344	1.055	0	0	1.000
Pyr(h)	0.498	0.139	3.574	0	0	1.000
Pyr(m)	1.656	1.639	1.010	0	0	1.000
Pyr(p)	1.418	1.268	1.118	0	0	1.000
Total PEP	2.896	2.583	1.121	0	0	1.000
PEP(c)	1.880	1.679	1.120	0	0	1.000
PEP(h)	1.536	1.672	0.919	0	1.000	0
G3P(h)	3.626	2.832	1.280	0	0	1.000
G3P(c)	1.837	0.966	1.901	0	0	1.000
Total ATP	1.967	1.609	1.223	0	0	1.000
ATP(h)	1.050	1.050	1.000	0	0	1.000
ATP(c)	6.691	6.234	1.073	0	0	1.000
ATP(m)	3.445	3.349	1.028	0	0	1.000
Total NADP	1.740	1.437	1.211	0	0	1.000
NADP(h)	1.506	1.447	1.041	0	0	1.000
NADP(c)	6.691	6.234	1.073	0	0	1.000
NADP(m)	3.445	3.349	1.028	0	0	1.000
Total NADPH	1.740	1.437	1.211	0	0	1.000
NADPH(h)	1.506	1.447	1.041	0	0	1.000
NADPH(c)	2.781	1.659	1.677	0	0	1.000
NADPH(m)	2.171	1.602	1.355	0	0	1.000
Total CO_2	0.839	0.070	11.998	0	0	1.000
CO_2(h)	0.001	0.002	0.788	0	1.000	0
CO_2(c)	5.434	4.581	1.186	0	0	1.000
CO_2(m)	2.451	2.172	1.128	0	0	1.000

Table S4.3. Activity of the CBC and starch metabolism in GC and M. The predicted mean flux values of reactions in the CBC cycle and starch metabolism are shown for G and M cells. The interpretation of the p-values is similar to that of Table S4.1. A horizontal bar indicates a failure of the test due to distributions consisting of a fixed value. Reaction names and index numbers in accord with AraCOREred.

Reaction [direction](Idx in AraCOREred)	Mean Flux G	Mean Flux M	Mean ratio (G/M)	Ho: $V_G = V_M$	Ho: $V_M > V_G$	Ho: $V_G > V_M$
CBC						
RuBisCO Carboxylation (6)	0.019	0.020	0.974	0	1.000	0
RuBisCO Oxygenation(85)	2.043e-06	2.043e-06	1.000	0	0	1.000
PGA kinase [Forward] (7)	0.495	0.278	1.779	0	0	1.000
GAP dehydrogenase (8)	0.864	0.959	0.901	0	1.000	0
TP isomerase [Forward] (9)	0.854	0.855	0.999	0	1.000	0
FBP aldolase [Forward] (10)	0	0	0	-	-	-
FBPase (11)	0	0	0	-	-	-
F6P transketolase (12)	0.007	0.007	0.974	0	1.000	0
SBP aldolase (13)	0	0	0	-	-	-
SBPase (14)	0	0	0	-	-	-
S7P transketolase (15)	0.006	0.007	0.974	0	1.000	0
Ru5P epimerase (16)	0.013	0.013	0.974	0	1.000	0
R5P isomerase (17)	0.006	0.007	0.973	0	1.000	0
Ru5P kinase (18)	0.019	0.020	0.974	0	1.000	0
PGA kinase [Backward] (7)	0	0	0	-	-	-
TP isomerase [Backward] (9)	0	0	0	-	-	-
FBP aldolase [Backward] (10)	0.004	0.145	0.025	0	1.000	0
Starch metabolism						
starch synthase (22)	2.949e-04	2.949e-04	1.000	0	0	1.000
starch synthase (23)	2.975e-04	3.068e-04	0.970	0	1.000	0
starch synthase (24)	2.610e-06	1.177e-05	0.222	0	1.000	0
amylase (26)	2.610e-06	2.107e-05	0.124	0	1.000	0
disproportionating enzyme (28)	0	9.165e-06	0	0	1.000	0
disproportionating enzyme (29)	0	9.298e-06	0	0	1.000	0

Table S4.4. A comparison of the maximum alternative optimal flux values of G and M cells for the reactions depicted in Figure 4.1. Results derived from the Flux Variability Analysis applied to the alternative optima space of RegrEx: $V_{max}G$, $V_{max}M$, the maximum flux value in the alternative optima space (as calculated by $RegrEx_{FVA}$) in G and M cells, respectively, and their difference are included. The p-values shown in this table correspond to a Mann-Whitney test comparing the distributions of flux values of G and M cells (V_G and V_M, respectively).

Idx. in Fig.1	Reaction Name (Idx in AraCOREred)	$V_{max}G$	$V_{max}M$	$V_{max}G$- $V_{max}M$	Ho: $V_G > V_M$	SubSystem
PSII	photosystem II (1)	0.125	0.125	0	0.841	light reactions
Cb6f	cytochrom b6f complex (2)	0.250	0.250	0	0.841	light reactions
PSI	photosystem I (3)	0.500	0.500	0	0.841	light reactions
ATPase	ATPase (5)	0.107	0.107	0	0.841	light reactions
1	PEP carboxylase (54)	0.466	0.054	0.412	1.000	gluconeogenesis
2	Cytosolic NADP-MDH (115)	0.168	0.105	0.063	1.000	pyruvate metabolism
3	Dicarboxylate transporter (339)	0.255	0.054	0.201	1.000	transport
4	Plastidial NADP-Malic Enzyme (113)	0.254	0.053	0.201	1.000	pyruvate metabolism
5	CO2 diffusion [Forward] (374)	1.000 e-06	0.012	-1.158e-02	0	transport
5	CO2 diffusion [Backward] (374)	0.234	0.033	0.201	1.000	transport
6	Import CO2 (413)	0.021	0.021	4.950e-05	0.006	import
7	Carbonic anhydrase (152)	0.466	0.054	0.412	1.000	carbon fixation
8	Glu synthase (FeS-Fd) (179)	0.250	0.026	0.224	1.000	glutamate synthesis
9	ferredoxin-NADP reductase (4)	0.235	0.250	-1.529e-02	0	light reactions
10	Export O2 (420)	0.020	0.020	1.310e-04	0.576	export
11	TP isomerase [Forward] (9)	0.874	0.924	-5.007e-02	0	Calvin-Benson cycle, glycolysis
12	TP/Pi translocator [Forward] (328)	0.888	1.000	-1.117e-01	0	transport
13	Di-/ri-carboxylate carrier [Forward] (346)	1.000	1.000	0	1.000	transport
13	Di-/ri-carboxylate carrier [Forward] (347)	1.000	1.000	0	1.000	transport
13	Di-/ri-carboxylate carrier [Forward] (348)	1.000	1.000	0	1.000	transport
14	Di-/ri-carboxylate carrier [Backward] (343)	1.000	1.000	0	1.000	transport
14	Di-/ri-carboxylate carrier [Backward] (344)	1.000	1.000	0	1.000	transport
14	Di-/ri-carboxylate carrier [Backward] (345)	1.000	1.000	0	1.000	transport
15	Mitochondrlial NAD-MDH [Backward] (80)	0.323	0.323	-4.698e-04	1.000	tricarboxylic acid cycle, glyoxylate cycle
15	Mitochondrial NADP-MDH (117)	0.769	0.746	0.023	1.000	pyruvate metabolism
16	FBP aldolase [Forward] (35)	0.034	0.218	-1.839e-01	0	sucrose synthesis, gluconeogenesis, glycolysis
17	FBPase (36)	0.090	0.006	0.083	1.000	sucrose synthesis, gluconeogenesis, glycolysis
17	PPi-dep. Phosphofructokinase [Backward] (136)	1.000 e-06	0.695	-6.950e-01	0	pyrophosphate recycling
18	G6P isomerase [Forward] (39)	0.034	0.218	-1.839e-01	0	sucrose synthesis, sucrose degradation, gluconeogenesis
19	Phosphoglucomutase [Forward] (40)	0.057	0.003	0.054	1.000	sucrose synthesis, sucrose degradation
20	TP/Pi translocator [Forward] (327)	1.000 e-06	1.000e-06	0	1.000	transport
20	TP/Pi translocator [Backward] (327)	0.386	0.600	-2.147e-01	0	transport

Table S4.5. A comparison of the maximum alternative optimal flux values of G and M cells for the reactions in the CBC and starch metabolism. Results derived from the Flux Variability Analysis applied to the alternative optima space of RegrEx: $V_{max}G$, $V_{max}M$, the maximum flux value in the alternative optima space (as calculated by $RegrEx_{FVA}$) in G and M cells, respectively, and their difference are included. The p-values shown in this table correspond to a Mann-Whitney test comparing the distributions of flux values of G and M cells (V_G and V_M, respectively).

Reaction [direction](Idx in GEM)	**$V_{max}G$**	**$V_{max}M$**	**$V_{max}G$- $V_{max}M$**	**Ho: $V_G > V_M$**
CBC				
RuBisCO Carboxylation (6)	0.021	0.023	-1.784e-03	0
RuBisCO Oxygenation(85)	2.514e-04	2.661e-04	-1.467e-05	-
PGA kinase [Forward] (7)	0.500	0.299	0.201	1.000
GAP dehydrogenase (8)	0.911	0.963	-5.196e-02	0
TP isomerase [Forward] (9)	0.874	0.924	-5.007e-02	0
FBP aldolase [Forward] (10)	0.005	1.000e-06	0.005	1.000
FBPase (11)	0.005	0.005	1.373e-04	1.96E-01
F6P transketolase (12)	0.007	0.008	-5.059e-04	0
SBP aldolase (13)	0.005	0.006	-8.402e-04	0
SBPase (14)	0.005	0.006	-8.402e-04	0
S7P transketolase (15)	0.007	0.008	-5.058e-04	0
Ru5P epimerase (16)	0.014	0.015	-1.012e-03	0
R5P isomerase (17)	0.007	0.007	-5.058e-04	0
Ru5P kinase (18)	0.021	0.023	-1.784e-03	0
PGA kinase [Backward] (7)	1.000e-06	1.000e-06	0	-
TP isomerase [Backward] (9)	1.000e-06	1.000e-06	0	0
FBP aldolase [Backward] (10)	0.020	0.204	-1.838e-01	0
Starch metabolism				
starch synthase (22)	0.020	0.204	-1.838e-01	1.000
starch synthase (23)	0.002	0.002	6.434e-04	0
starch synthase (24)	0.005	0.006	-7.696e-04	0
amylase (26)	0.004	0.003	0.001	0
disproportionating enzyme (28)	0.004	0.003	0.001	0
disproportionating enzyme (29)	0.004	0.003	0.001	0
disproportionating enzyme (30)	0.004	0.003	0.001	1.76E-05
(starch) phosphorylase (32)	0.002	0.002	1.543e-04	1.000

Table S4.6. A comparison of the predicted metabolic state of G and M cells after imposing additional experimental constraints. This table shows the analogous results, presented in Table S4.1, when additional constraints are taken in consideration. Concretely, the carboxylation to oxygenation ratio of RubisCO is constrained to stay within 1.5 and 4. Additionally, the flux through the reactions in the CBC: the *sedoheptulose 1,7-bisphosphate aldolase* and the *sedoheptulose-1,7-bisphosphatase* is constrained to carry a positive flux (Details in Materials & Methods, section 4.4). A horizontal bar indicates a failure of the test due to distributions consisting of a fixed value. Reaction names and index numbers in accord with AraCOREred.

Idx in Fig 1	Reaction Name (Idx in AraCOREred)	Mean Flux G	Mean Flux M	Mean ratio (G/M)	Ho: $V_G = V_M$	Ho: $V_M > V_G$	Ho: $V_G > V_M$	SubSystem
PSII	photosystem II (1)	0.125	0.125	1.000	-	-	-	light reactions
Cb6f	cytochrom b6f complex (2)	0.250	0.250	1.000	-	-	-	light reactions
PSI	photosystem I (3)	0.500	0.500	1.000	-	-	-	light reactions
ATPase	ATPase (5)	0.107	0.107	1.000	-	-	-	light reactions
1	PEP carboxylase (54)	0.420	0.011	37.190	0	0	1.000	gluconeogenesis
2	Mal dehydrogenase (115)	0.161	0.086	1.866	0	0	1.000	pyruvate metabolism
3	Dicarboxylate transporter (339)	0.223	0.011	19.971	0	0	1.000	transport
4	Mal dehydrogenase (113)	0.221	0.009	24.697	0	0	1.000	pyruvate metabolism
5	CO2 diffusion [Forward] (374)	0	0.013	0	0	1.000	0	transport
5	CO2 diffusion [Backward] (374)	0.199	7.085e-08	2.812e+06	0	0	1.000	transport
6	Import CO2 (413)	0.021	0.021	1.000	0	1.000	0	import
7	HCO3 dehydratase (152)	0.420	0.011	37.190	0	0	1.000	carbon fixation
8	Glu synthase (FeS-Fd) (179)	0.120	1.670e-04	719.020	0	0	1.000	glutamate synthesis
9	ferredoxin-NADP reductase (4)	0.130	0.250	0.520	0	1.000	0	light reactions
10	Export O2 (420)	0.020	0.020	1.000	0	1.000	0	export
11	TP isomerase [Forward] (9)	0.854	0.825	1.035	0	0	1.000	Calvin-Benson cycle, glycolysis
12	TP/Pi translocator [Forward] (328)	0.856	1.000	0.856	0	1.000	0	transport
13	Di-/ri-carboxylate carrier [Forward] (346)	0.362	0.335	1.079	0	0	1.000	transport
13	Di-/ri-carboxylate carrier [Backward] (346)	0.355	0.335	1.060	0	0	1.000	transport
13	Di-/ri-carboxylate carrier [Forward] (347)	0.362	0.338	1.069	0	0	1.000	transport
13	Di-/ri-carboxylate carrier [Backward] (347)	0.332	0.309	1.076	0	0	1.000	transport
13	Di-/ri-carboxylate carrier [Forward] (348)	0.327	0.310	1.056	1.455e-05	7.273e-06	1.000	transport
13	Di-/ri-carboxylate carrier [Backward] (348)	0.332	0.305	1.089	0	0	1.000	transport
14	Di-/ri-carboxylate carrier [Forward] (343)	0.249	0.208	1.199	0	0	1.000	transport
14	Di-/ri-carboxylate carrier [Backward] (343)	0.753	0.720	1.046	0	0	1.000	transport
14	Di-/ri-carboxylate carrier [Forward] (344)	0.022	0.194	0.113	0	1.000	0	transport
14	Di-/ri-carboxylate carrier [Backward] (344)	0.058	0	Inf	0	0	1.000	transport
14	Di-/ri-carboxylate carrier [Forward] (345)	0	0.693	0	0	1.000	0	transport
14	Di-/ri-carboxylate carrier [Backward] (345)	0.022	0.194	0.113	0	1.000	0	transport
15	Mal dehydrogenase [Backward] (80)	0.036	7.590e-04	47.157	0	0	1.000	tricarboxylic acid cycle, glyoxylate cycle
15	Mal dehydrogenase (117)	0	0	0	-	-	-	pyruvate metabolism
16	FBP aldolase [Forward] (35)	0.394	0.405	0.973	0	1.000	0	sucrose synthesis, gluconeogenesis, glycolysis
17	FBPase (36)	0.125	0.125	1.000	-	-	-	sucrose synthesis, gluconeogenesis, glycolysis
17	PPi-dep. Phosphofructokinase [Backward] (136)	0.250	0.250	1.000	-	-	-	pyrophosphate recycling
18	G6P isomerase [Forward] (39)	0.500	0.500	1.000	-	-	-	sucrose synthesis, sucrose degradation, gluconeogenesis
19	Phosphoglucomutase [Forward] (40)	0.107	0.107	1.000	-	-	-	sucrose synthesis, sucrose degradation
20	TP/Pi translocator [Forward] (327)	0.420	0.011	37.190	0	0	1.000	transport
20	TP/Pi translocator [Backward] (327)	0.161	0.086	1.866	0	0	1.000	transport

Table S4.7. Activity of the CBC and starch metabolism in G and M cells after imposing additional experimental constraints. This table shows the analogous results to that of Table S4.3, when additional constraints are taken in consideration. Concretely, the carboxylation to oxygenation ratio of RubisCO is constrained to stay within 1.5 and 4. Additionally, the flux through the reactions in the CBC: the *sedoheptulose 1,7-bisphosphate aldolase* and the *sedoheptulose-1,7-bisphosphatase* is constrained to carry a positive flux (details in Materials & Methods, section 4.4). A horizontal bar indicates a failure of the test due to distributions consisting of a fixed value. Reaction names and index numbers in accord with AraCOREred.

Reaction [direction](Idx in GEM)	**Mean Flux G**	**Mean Flux M**	**Mean ratio (G/M)**	**Ho: $V_G = V_M$**	**Ho: $V_M > V_G$**	**Ho: $V_G > V_M$**
CBC						
RuBisCO Carboxylation (6)	0.022	0.022	1.000	0	1.000	0
RuBisCO Oxygenation(85)	0.005	0.005	1.000	0	1.000	0
PGA kinase [Forward] (7)	0.467	0.255	1.832	0	0	1.000
GAP dehydrogenase (8)	0.864	0.959	0.901	0	1.000	0
TP isomerase [Forward] (9)	0.854	0.825	1.035	0	0	1.000
FBP aldolase [Forward] (10)	5.544e-04	0	Inf	0	0	1.000
FBPase (11)	0.001	0.001	1.000	-	-	-
F6P transketolase (12)	0.009	0.009	1.000	0	1.000	0
SBP aldolase (13)	0.001	0.001	1.000	-	-	-
SBPase (14)	0.001	0.001	1.000	-	-	-
S7P transketolase (15)	0.009	0.009	1.000	0	1.000	0
Ru5P epimerase (16)	0.018	0.018	1.000	0	1.000	0
R5P isomerase (17)	0.009	0.009	1.000	0	1.000	0
Ru5P kinase (18)	0.027	0.027	1.000	0	1.000	0
PGA kinase [Backward] (7)	0	0	0	-	-	-
TP isomerase [Backward] (9)	0	0	0	-	-	-
FBP aldolase [Backward] (10)	0.004	0.176	0.021	0	1.000	0
Starch metabolism						
starch synthase (22)	2.949e-04	2.949e-04	1.000	0	0	1.000
starch synthase (23)	2.975e-04	3.065e-04	0.971	0	1.000	0
starch synthase (24)	2.632e-06	1.203e-05	0.219	0	1.000	0
amylase (26)	2.632e-06	2.106e-05	0.125	0	1.000	0
disproportionating enzyme (28)	2.239e-08	9.423e-06	0.002	0	1.000	0
disproportionating enzyme (29)	0	9.031e-06	0	0	1.000	0

Table S4.8. Predicted Flux-Sums of selected metabolites in the AraCOREred model after imposing additional experimental constraints. This table presents the analogous results of Table S4.2 when additional constraints are taken in consideration. Concretely, the carboxylation to oxygenation ratio of RubisCO is constrained to stay within 1.5 and 4. Additionally, the flux through the reactions in the CBC: the *sedoheptulose 1,7-bisphosphate aldolase* and the *sedoheptulose-1,7-bisphosphatase* is constrained to carry a positive flux (details in Materials & Methods, section 4.4).

Metabolite (compartment)	Mean FluxSum G	Mean FluxSum M	Ratio (G/M)	Ho: $V_G = V_M$	Ho: $V_M > V_G$	Ho: $V_G > V_M$
Total Mal	5.615	4.722	1.189	0	0	1.000
Mal(c)	2.678	2.359	1.135	0	0	1.000
Mal(m)	2.355	2.186	1.077	0	0	1.000
Mal(p)	0.006	0.006	1.001	0	1.000	0
Mal(h)	2.155	1.443	1.493	0	0	1.000
Total Suc	0.070	5.574e-05	1.258e+03	0	0	1.000
Futile Cycle Suc	0.070	1.203e-05	5.822e+03	0	1.000	0
Total OAA	7.570	6.777	1.117	0	0	1.000
OAA(c)	4.191	3.677	1.140	0	0	1.000
OAA(m)	2.156	2.017	1.069	0	0	1.000
OAA(p)	1.326	1.494	0.888	0	1.000	0
OAA(h)	1.713	1.425	1.202	0	0	1.000
Total Pyr	3.510	2.980	1.178	0	0	1.000
Pyr(c)	1.413	1.303	1.084	0	0	1.000
Pyr(h)	0.442	0.057	7.725	0	0	1.000
Pyr(m)	1.656	1.639	1.010	0	0	1.000
Pyr(p)	1.413	1.264	1.117	0	0	1.000
Total PEP	2.926	2.542	1.151	0	0	1.000
PEP(c)	1.937	1.679	1.154	0	0	1.000
PEP(h)	1.538	1.672	0.920	0	1.000	0
G3P(h)	3.348	2.539	1.318	0	0	1.000
G3P(c)	1.693	0.845	2.003	0	0	1.000
Total ATP	1.802	1.417	1.271	0	0	1.000
ATP(h)	0.968	0.968	1.000	0	0	1.000
ATP(c)	6.691	6.226	1.075	0	0	1.000
ATP(m)	3.445	3.344	1.030	0	0	1.000
Total NADP	1.740	1.442	1.207	0	0	1.000
NADP(h)	1.506	1.440	1.046	0	0	1.000
NADP(c)	6.691	6.226	1.075	0	0	1.000
NADP(m)	3.445	3.344	1.030	0	0	1.000
Total NADPH	1.740	1.442	1.207	0	0	1.000
NADPH(h)	1.506	1.440	1.046	0	0	1.000
NADPH(c)	2.759	1.619	1.704	0	0	1.000
NADPH(m)	2.116	1.580	1.339	0	0	1.000
Total CO_2	0.839	0.048	17.473	0	0	1.000
CO_2(h)	0.007	0.007	1.000	0	1.000	0
CO_2(c)	5.615	4.722	1.189	0	0	1.000
CO_2(m)	2.678	2.359	1.135	0	0	1.000

Table S4.9. Redistribution of the percentage of ^{13}C label enrichment in primary metabolites. M and G cells were fed with ^{13}-$NaHCO_3$ and harvested after 30 min and 60 min in the light. Values in bold and underline type are significantly different between M and G cells according to Students *t*-test (P <0.05) in the same time point. Data presented are mean ± SE (n = 3).

		% of 13C enrichment							
Metabolite	**m/z**	**M30**	**SE**	**M60**	**SE**	**GC30**	**SE**	**GC60**	**SE**
Gly	**102**	0.010	0.004	0.005	0.001	0.015	0.003	0.006	0.002
Ser	**306**	0.016	0.014	0.009	0.005	0.007	0.004	**0.036**	**0.004**
Ser	**204**	0.101	0.052	0.060	0.020	0.111	0.036	0.097	0.007
Homoserine	**128**	0.449	0.272	0.380	0.028	0.282	0.124	0.332	0.237
Glycolate	**205**	0.008	0.008	0.007	0.007	0.015	0.008	0.017	0.008
Val	**218**	0.240	0.180	0.095	0.091	0.051	0.027	0.162	0.035
Ala	**188**	0.416	0.198	0.361	0.141	1.476	1.260	0.459	0.355
Thr	**219**	0.458	0.238	0.342	0.132	0.795	0.330	0.634	0.067
Thr	**291**	0.927	0.438	0.748	0.260	1.489	0.447	**1.361**	**0.118**
Pro	**142**	0.197	0.100	0.164	0.021	0.321	0.151	**0.367**	**0.012**
Asp	**218**	0.018	0.018	0.012	0.012	**1.638**	**0.947**	0.786	0.622
Asp	**232**	0.045	0.004	0.037	0.002	**0.078**	**0.005**	0.038	0.001
Leu	**158**	0.034	0.000	0.033	0.000	0.031	0.002	0.034	0.002
Ile	**218**	0.038	0.038	0.091	0.091	0.068	0.049	0.203	0.101
Glu	**156**	0.010	0.001	0.014	0.004	0.020	0.009	0.022	0.007
Glycerate	**292**	1.676	0.636	1.922	0.653	**3.521**	**0.069**	3.609	0.746
Lactate	**117**	0.111	0.075	0.421	0.033	0.344	0.172	0.268	0.163
Glycerol	**293**	0.072	0.041	0.081	0.043	0.075	0.055	0.114	0.018
Succ	**172**	0.176	0.080	0.283	0.100	**0.752**	**0.154**	**0.506**	**0.044**
Succ	**247**	0.219	0.071	0.266	0.102	**0.954**	**0.071**	**0.607**	**0.039**
Mal	**233**	0.040	0.001	0.053	0.010	**0.100**	**0.001**	0.041	0.002
Fum	**245**	0.098	0.026	0.084	0.014	0.094	0.033	0.128	0.019
GABA	**174**	0.015	0.003	0.024	0.008	nd	nd	nd	nd
Erythritol	**217**	0.425	0.130	0.438	0.149	2.212	1.808	1.207	0.871
Sucrose	**437**	0.000015	0.000007	0.000018	0.000009	0.000104	0.000068	0.000036	0.000010

Table S4.10. Total ^{13}C-enrichment in primary metabolites. Experimental and statistical analysis as described in the Table S4.9.

	Total 13C enrichment								
Metabolite	**m/z**	**M30**	**SE**	**M60**	**SE**	**GC30**	**SE**	**GC60**	**SE**
Gly	**102**	3257	1665	4837	624	1083	496	4738	997
Ser	**306**	71.7	22.1	194.9	40.0	135.6	90.9	**303.9**	**23.4**
Ser	**204**	1642	976	1786	579	921	332	836	103
Homoserine	**128**	1524	976	628	44	1331	526	2202	1073
Glycolate	**205**	658	658	646	646	1212	609	1370	595
Val	**218**	30.3	7.2	20.9	13.7	**7.3**	**4.7**	13.2	0.9
Ala	**188**	694	513	715	335	479	228	69	36
Thr	**219**	276	105	270	75	136	38	128	13
Thr	**291**	735	253	740	201	400	104	377	49
Pro	**142**	251	92	186	24	133	50	**52**	**15**
Asp	**218**	12.0	12.0	12.5	12.5	14.8	1.7	8.9	4.7
Asp	**232**	1533	76	1303	37	**2352**	**133**	1154	6
Leu	**158**	499	23	514	29	**386**	**31**	437	33
Ile	**218**	1.84	1.84	8.88	8.88	2.72	1.36	17.59	7.21
Glu	**156**	6938	109	6594	2023	**2249**	**566**	2423	313
Glycerate	**292**	92.6	37.0	78.0	15.7	39.9	0.3	47.9	8.9
Lactate	**117**	356	251	148	13	349	130	500	230
Glycerol	**293**	3363	1273	2824	992	5972	2954	1522	283
Succ	**172**	5380	3759	1361	336	779	246	905	165
Succ	**247**	303	88	226	64	**87**	**12**	134	22
Mal	**233**	1341	31	1790	330	**3242**	**41**	1289	53
Fum	**245**	713	241	651	113	548	179	529	19
GABA	**174**	14330	2650	9640	3556	nd	nd	nd	nd
Erythritol	**217**	206	53	186	43	133	64	**60**	**15**
Sucrose	**437**	477177	82661	607521	270298	427991	174100	486175	190880

Table S4.11. ^{13}C-enrichment in primary metabolites. Experimental and statistical analysis as described in Table S4.9.

	13C enrichment								
Metabolite	***m/z***	**M30**	**SE**	**M60**	**SE**	**GC30**	**SE**	**GC60**	**SE**
Gly 2TMS	**102**	4.7	0.2	5.1	0.1	3.6	0.7	5.0	0.2
Ser	**306**	0.8	0.5	1.1	0.2	1.0	0.6	**3.3**	**0.1**
Ser	**204**	8.9	0.6	9.2	0.4	8.9	0.5	8.9	0.2
Homoserine	**128**	16.1	0.3	15.4	0.1	15.9	0.5	**16.5**	**0.0**
Glycolate	**205**	2.3	2.3	2.1	2.1	4.3	2.1	4.8	2.1
Val	**218**	2.5	1.3	1.3	1.1	0.6	0.3	1.5	0.2
Ala	**188**	9.4	1.5	12.1	4.2	12.8	1.4	5.1	2.9
Thr	**219**	8.9	0.2	8.5	0.1	9.1	0.4	**8.9**	**0.0**
Thr	**291**	21.6	0.5	21.2	0.3	22.5	0.1	22.4	0.5
Pro	**142**	5.6	0.1	5.4	0.0	**5.3**	**0.0**	4.3	0.6
Asp	**218**	0.5	0.47	0.4	0.39	**4.4**	**1.4**	2.3	1.3
Asp	**232**	8.3	0.6	7.0	0.3	**13.5**	**0.8**	6.6	0.1
Leu	**158**	4.1	0.1	4.1	0.1	**3.5**	**0.3**	3.9	0.2
Ile	**218**	0.3	0.3	0.9	0.9	0.4	0.2	1.9	0.9
Glu	**156**	6.3	1.6	7.8	0.4	5.1	2.4	**5.0**	**1.0**
Glycerate	**292**	10.5	0.3	11.4	0.6	**11.9**	**0.1**	12.6	0.1
Lactate	**117**	5.3	2.6	7.9	0.6	8.8	0.2	8.1	0.2
Uracil	**241**	5.2	0.2	5.6	0.5	**7.1**	**0.0**	6.4	0.2
Uracil	**255**	9.1	0.1	9.3	0.1	**10.8**	**0.4**	**10.8**	**0.1**
Adipic acid	**111**	1.9	0.1	1.9	0.1	1.8	0.0	2.1	0.1
Adipic acid	**141**	6.0	0.6	5.9	0.2	6.0	0.3	6.5	0.3
Threonate	**292**	10.0	0.2	10.1	0.4	6.5	3.7	nd	nd
Salicylic acid	**267**	8.8	1.2	9.3	1.1	nd	nd	nd	nd
Glycerol	**293**	12.1	0.6	12.0	0.5	12.4	0.1	12.8	0.2
OXA	**175**	nd	nd	nd	nd	14.5	0.3	nd	nd
Succ	**172**	22.6	6.1	17.8	0.5	22.5	1.7	**21.0**	**1.1**
Succ	**247**	7.3	0.5	6.9	0.3	**9.0**	**0.3**	**9.0**	**1.0**
Mal	**233**	7.3	0.0	9.7	1.8	**18.0**	**0.1**	7.3	0.3
Citramalate	**247**	0.3	0.28	0.1	0.097	0.8	0.4	**1.3**	**0.8**
Fum	**245**	7.6	0.4	7.2	0.0	6.7	1.6	8.2	0.5
GABA	**174**	6.6	0.1	6.4	0.1	nd	nd	nd	nd
Erythritol	**217**	8.6	0.2	8.3	0.2	9.3	1.0	6.6	1.8
myo inositol	**191**	4.9	4.9	4.8	4.8	7.2	4.1	11.6	2.1
Sucrose	**437**	2.5	0.8	2.5	0.1	**4.7**	**0.3**	**3.7**	**0.4**
Trehalose	**169**	3.9	2.2	5.6	2.8	5.1	2.7	5.0	2.5
Maltose	**361**	7.6	3.8	8.7	2.2	nd	nd	8.1	0.3
Isomaltose	**361**	nd	nd	12.6	0.6	nd	nd	nd	nd

Table S4.12. Content of the metabolites analyzed in this study. The content (ng) of each metabolite was obtained using the equation of a linear regression of the peak area obtained from different concentrations of standard compounds.

	Concentration (ng)							
Metabolite	**M30**	**SE**	**M60**	**SE**	**GC30**	**SE**	**GC60**	**SE**
Gly	675.1	319.4	949.3	103.7	266.1	84.7	959.5	238.7
Ser	171.7	94.6	189.2	55.6	100.6	32.8	93.5	9.3
Homoserine	92.3	57.6	40.9	2.8	81.7	30.6	133.6	65.2
Glycolate	291.5	7.3	290.9	6.1	283.0	4.0	284.3	2.8
Val	17.1	4.7	29.4	8.4	9.6	2.1	9.3	0.7
Ala	71.7	54.4	58.7	32.7	42.1	22.1	23.4	8.4
Thr	34.5	12.5	34.8	9.3	17.8	4.7	16.8	1.8
Pro	44.2	15.9	34.3	4.3	24.8	9.4	**11.7**	**2.0**
Asp	184.4	3.9	187.1	3.9	**174.1**	**0.9**	**175.0**	**1.5**
Leu	120.6	2.3	124.9	3.7	111.3	0.8	112.8	1.5
Ile	17.2	6.2	22.5	6.3	9.9	2.8	10.0	0.7
Glu	673.4	207.5	697.1	263.6	224.6	92.0	324.1	146.4
Glycerate	8.9	3.7	7.0	1.7	3.4	0.0	3.8	0.7
Lactate	46.4	29.2	18.7	0.9	40.1	15.5	60.7	27.5
Salicylic acid	10.7	7.6	5.3	2.2	0.5	0.1	0.5	0.1
Glycerol	289.4	118.7	243.6	93.2	482.3	239.2	118.2	20.1
Succ	192.6	93.8	77.4	20.6	33.3	8.4	42.5	5.7
Mal	182.9	4.3	184.3	1.1	179.6	1.9	177.7	0.1
Fum	94.9	33.1	90.4	15.7	81.0	22.4	**65.8**	**7.0**
GABA	489.1	97.1	343.7	120.0	**135.4**	**78.2**	**35.2**	**20.3**
Erythritol	24.1	6.6	22.7	5.5	15.8	8.0	11.5	4.7
Sucrose	246540	86181	239692	101307	97619	43785	124067	39544

Bibliography

Agren, R., Bordel, S., Mardinoglu, A., Pornputtapong, N., Nookaew, I., & Nielsen, J. (2012). Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. *PLoS Computational Biology*, *8*(5).

Agren, R., Mardinoglu, A., Asplund, A., Kampf, C., Uhlen, M., & Nielsen, J. (2014). Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. *Molecular Systems Biology*, *10*, 721. https://doi.org/10.1002/msb.145122

Alonso, A. P., Raymond, P., Rolin, D., & Dieuaide-Noubhani, M. (2007). Substrate cycles in the central metabolism of maize root tips under hypoxia. *Phytochemistry*, *68*(16–18), 2222–2231. https://doi.org/10.1016/j.phytochem.2007.04.022

Alonso, A. P., Vigeolas, H., Raymond, P., Rolin, D., & Dieuaide-noubhani, M. (2005). A New Substrate Cycle in Plants. Evidence for a High Glucose-Phosphate-to-Glucose Turnover from in Vivo Steady-State and Pulse-Labeling Experiments with [13C]Glucose and [14C]Glucose. *Plant Physiol*, *138*(August), 2220–2232. https://doi.org/10.1104/pp.105.062083.found

Antunes, W. C., Provart, N. J., Williams, T. C. R., & Loureiro, M. E. (2012). Changes in stomatal function and water use efficiency in potato plants with altered sucrolytic activity. *Plant, Cell and Environment*, *35*(4), 747–759. https://doi.org/10.1111/j.1365-3040.2011.02448.x

Araújo, W. L., Nunes-Nesi, A., Osorio, S., Usadel, B., Fuentes, D., Nagy, R., … Fernie, A. R. (2011). Antisense inhibition of the iron-sulphur subunit of succinate dehydrogenase enhances photosynthesis and growth in tomato via an organic acid-mediated effect on stomatal aperture. *The Plant Cell*, *23*(2), 600–27. https://doi.org/10.1105/tpc.110.081224

Arnold, A., & Nikoloski, Z. (2014). Bottom-up Metabolic Reconstruction of Arabidopsis and Its Application to Determining the Metabolic Costs of Enzyme Production. *Plant Physiology*, *165*, 1380–1391. https://doi.org/10.1104/pp.114.235358

Aubry, S., Aresheva, O., Reyna-Llorens, I., Smith-Unna, R. D., Hibberd, J. M., & Genty, B. (2016). A Specific Transcriptome Signature for Guard Cells from the C4 Plant Gynandropsis gynandra. *Plant Physiology*, *170*(3), 1345–57. https://doi.org/10.1104/pp.15.01203

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., … Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Azoulay-Shemer, T., Palomares, A., Bagheri, A., Israelsson-Nordstrom, M., Engineer, C. B., Bargmann, B. O. R., … Schroeder, J. I. (2015). Guard cell photosynthesis is critical for stomatal turgor production, yet does not directly mediate CO<inf>2</inf>- and ABA-induced stomatal closing. *Plant Journal*, *83*(4), 567–581. https://doi.org/10.1111/tpj.12916

Baart, G. J. E., & Martens, D. E. (2012). Genome-Scale Metabolic Models: Reconstruction and Analysis (pp. 107–126). https://doi.org/10.1007/978-1-61779-346-2_7

Bar-Even, A., Flamholz, A., Noor, E., & Milo, R. (2012). Rethinking glycolysis: on the biochemical logic of metabolic pathways. *Nature Chemical Biology*, *8*(6), 509–517. https://doi.org/10.1038/nchembio.971

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., … Soboleva, A. (2013). NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Research*, *41*(Database issue), D991-5. https://doi.org/10.1093/nar/gks1193

Bates, G. W., Rosenthal, D. M., Sun, J., Chattopadhyay, M., Peffer, E., Yang, J., … Jones, A. M. (2012).

A Comparative Study of the Arabidopsis thaliana Guard-Cell Transcriptome and Its Modulation by Sucrose. *PLoS ONE*, *7*. https://doi.org/10.1371/journal.pone.0049641

Bauer, H., Ache, P., Lautner, S., Fromm, J., Hartung, W., Al-Rasheid, K. A. S., … Hedrich, R. (2013). The stomatal response to reduced relative humidity requires guard cell-autonomous ABA synthesis. *Current Biology*, *23*, 53–57. https://doi.org/10.1016/j.cub.2012.11.022

Becker, S. A., & Palsson, B. O. (2008). Context-specific metabolic networks are consistent with experiments. *PLoS Computational Biology*, *4*(5).

Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., & Huala, E. (2015). The arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. *Genesis*, *53*(8), 474–485. https://doi.org/10.1002/dvg.22877

Binns, M., de Atauri, P., Vlysidis, A., Cascante, M., & Theodoropoulos, C. (2015). Sampling with poling-based flux balance analysis: optimal versus sub-optimal flux space analysis of Actinobacillus succinogenes. *BMC Bioinformatics*, *16*(1), 49. https://doi.org/10.1186/s12859-015-0476-5

BioMart (Ensembl). (n.d.). Retrieved from http://www.ensembl.org/biomart/martview/2a3c1aa45a4126aa9947f83d577eee2b

Blazier, A. S., & Papin, J. A. (2012). Integration of expression data in genome-scale metabolic network reconstructions. *Frontiers in Physiology*. https://doi.org/10.3389/fphys.2012.00299

Booker, F., Burkey, K., Morgan, P., Fiscus, E., & Jones, A. (2012). Minimal influence of G-protein null mutations on ozone-induced changes in gene expression, foliar injury, gas exchange and peroxidase activity in Arabidopsis thaliana L. *Plant, Cell & Environment*, *35*(4), 668–81. https://doi.org/10.1111/j.1365-3040.2011.02443.x

Bordbar, A., Monk, J. M., King, Z. a, & Palsson, B. O. (2014). Constraint-based models predict metabolic and associated cellular functions. *Nature Reviews. Genetics*, *15*, 107–20. https://doi.org/10.1038/nrg3643

Boyd, S., & Vandenberghe, L. (2010). *Convex Optimization*. *Optimization Methods and Software* (Vol. 25). https://doi.org/10.1080/10556781003625177

Briggs, G. E., & Haldane, J. B. (1925). A Note on the Kinetics of Enzyme Action. *The Biochemical Journal*, *19*(2), 338–9. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16743508

Bruce A. Murtagh, M. A. S. (n.d.). Minos User's Manual. *Systems Optimization Laboratory. Department of Operations Research. Stanford University*.

Butte, A. (2002). The use and analysis of microarray data. *Nature Reviews Drug Discovery*, *1*(12), 951–960. https://doi.org/10.1038/nrd961

Candes, E., & Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, *35*, 2313–2351. https://doi.org/10.1214/009053606000001523

Caspi, R., Billington, R., Ferrer, L., Foerster, H., Fulcher, C. A., Keseler, I. M., … Karp, P. D. (2016). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research*, *44*(D1), D471–D480. https://doi.org/10.1093/nar/gkv1164

Chai, L. E., Loh, S. K., Low, S. T., Mohamad, M. S., Deris, S., & Zakaria, Z. (2014). A review on the computational approaches for gene regulatory network construction. *Computers in Biology and Medicine*, *48*, 55–65. https://doi.org/10.1016/j.compbiomed.2014.02.011

Chen, C., Grennan, K., Badner, J., Zhang, D., Gershon, E., Jin, L., & Liu, C. (2011). Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods. *PLoS ONE*, *6*(2), e17238. https://doi.org/10.1371/journal.pone.0017238

Chen, Z.-H., Hills, A., Batz, U., Amtmann, A., Lew, V. L., & Blatt, M. R. (2012). Systems Dynamic

Modeling of the Stomatal Guard Cell Predicts Emergent Behaviors in Transport, Signaling, and Volume Control. *Plant Physiology*, *159*(3), 1235–1251. https://doi.org/10.1104/pp.112.197350

Chung, B., Lee, D.-Y., Llaneras, F., Pico, J., Strogatz, S., Segre, D., … Lee, S. (2009). Flux-sum analysis: a metabolite-centric approach for understanding the metabolic network. *BMC Systems Biology*, *3*(1), 117. https://doi.org/10.1186/1752-0509-3-117

Colijn, C., Brandes, A., Zucker, J., Lun, D. S., Weiner, B., Farhat, M. R., … Galagan, J. E. (2009). Interpreting expression data with metabolic flux models: Predicting Mycobacterium tuberculosis mycolic acid production. *PLoS Computational Biology*, *5*. https://doi.org/10.1371/journal.pcbi.1000489

Collins, S. B., Reznik, E., & Segrè, D. (2012). Temporal Expression-based Analysis of Metabolism. *PLoS Computational Biology*, *8*. https://doi.org/10.1371/journal.pcbi.1002781

Cook, D. J., & Nielsen, J. (2017). Genome-scale metabolic models applied to human health and disease. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, e1393. https://doi.org/10.1002/wsbm.1393

Covert, M. W. (2002). Transcriptional Regulation in Constraints-based Metabolic Models of Escherichia coli. *Journal of Biological Chemistry*, *277*(31), 28058–28064. https://doi.org/10.1074/jbc.M201691200

Covert, M. W., Schilling, C. H., & Palsson, B. (2001). Regulation of gene expression in flux balance models of metabolism. *Journal of Theoretical Biology*, *213*(1), 73–88. https://doi.org/10.1006/jtbi.2001.2405

Dal'Molin, C. G. de O., Quek, L.-E., Palfreyman, R. W., Brumbley, S. M., & Nielsen, L. K. (2010). C4GEM, a genome-scale metabolic model to study C4 plant metabolism. *Plant Physiology*, *154*(4), 1871–1885.

Daloso, D. M., Antunes, W. C., Pinheiro, D. P., Waquim, J. P., Ara??jo, W. L., Loureiro, M. E., … Williams, T. C. R. (2015). Tobacco guard cells fix CO2 by both Rubisco and PEPcase while sucrose acts as a substrate during light-induced stomatal opening. *Plant, Cell and Environment*, *38*(11), 2353–2371. https://doi.org/10.1111/pce.12555

Daloso, D. M., dos Anjos, L., & Fernie, A. R. (2016). Roles of sucrose in guard cell regulation. *New Phytologist*. https://doi.org/10.1111/nph.13950

Daloso, D. M., Müller, K., Obata, T., Florian, A., Tohge, T., Bottcher, A., … Fernie, A. R. (2015). Thioredoxin, a master regulator of the tricarboxylic acid cycle in plant mitochondria. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(11), E1392-400. https://doi.org/10.1073/pnas.1424840112

Daloso, D. M., Williams, T. C. R., Antunes, W. C., Pinheiro, D. P., Müller, C., Loureiro, M. E., & Fernie, A. R. (2016). Guard cell-specific upregulation of sucrose synthase 3 reveals that the role of sucrose in stomatal function is primarily energetic. *New Phytologist*, *209*(4), 1470–1483. https://doi.org/10.1111/nph.13704

Dancer, J., Hatzfeld, W. D., & Stitt, M. (1990). Cytosolic cycles regulate the turnover of sucrose in heterotrophic cell-suspension cultures of Chenopodium rubrum L. *Planta*, *182*(2), 223–231. https://doi.org/10.1007/BF00197115

De Martino, D., Mori, M., & Parisi, V. (2015). Uniform sampling of steady states in metabolic networks: heterogeneous scales and rounding. *PloS One*, *10*(4), e0122670. https://doi.org/10.1371/journal.pone.0122670

de Oliveira Dal'Molin, C. G., Quek, L.-E., Palfreyman, R. W., Brumbley, S. M., & Nielsen, L. K. (2010). AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis. *Plant Physiology*, *152*, 579–589. https://doi.org/10.1104/pp.109.148817

Dias, O., Rocha, M., Ferreira, E. C., & Rocha, I. (2015). Reconstructing genome-scale metabolic models

with merlin. *Nucleic Acids Research*, *43*(8), 3899–3910. https://doi.org/10.1093/nar/gkv294

Dielman, T. E. (2005). Least absolute value regression: recent contributions. *Journal of Statistical Computation and Simulation*, *75*(4), 263–286. https://doi.org/10.1080/00949650420002236 80

Dowell, R. D., Ryan, O., Jansen, A., Cheung, D., Agarwala, S., Danford, T., … Boone, C. (2010). Genotype to Phenotype: A Complex Problem. *Science*, *328*(5977).

Duarte, N. C., Becker, S. A., Jamshidi, N., Thiele, I., Mo, M. L., Vo, T. D., … Palsson, B. Ø. (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 1777–1782. https://doi.org/10.1073/pnas.0610772104

Eastmond, P. J., Astley, H. M., Parsley, K., Aubry, S., Williams, B. P., Menard, G. N., … Hibberd, J. M. (2015). Arabidopsis uses two gluconeogenic gateways for organic acids to fuel seedling establishment. *Nature Communications*, *6*, 6659. https://doi.org/10.1038/ncomms7659

Edwards, J. S., & Palsson, B. O. (1999). Systems properties of the Haemophilus influenzae Rd metabolic genotype. *The Journal of Biological Chemistry*, *274*(25), 17410–6. https://doi.org/10.1074/JBC.274.25.17410

Faria, J. P., Overbeek, R., Xia, F., Rocha, M., Rocha, I., & Henry, C. S. (2013). Genome-scale bacterial transcriptional regulatory networks: reconstruction and integrated analysis with metabolic models. *Briefings in Bioinformatics*, bbs071. https://doi.org/10.1093/bib/bbs071

Feist, A. M., & Palsson, B. O. (2010). The biomass objective function. *Current Opinion in Microbiology*, *13*(3), 344–9. https://doi.org/10.1016/j.mib.2010.03.003

Fernie, A. R., & Martinoia, E. (2009). Malate. Jack of all trades or master of a few? *Phytochemistry*, *70*(7), 828–32. https://doi.org/10.1016/j.phytochem.2009.04.023

Fettke, J., & Fernie, A. R. (2015). Intracellular and cell-to-apoplast compartmentation of carbohydrate metabolism. *Trends in Plant Science*, *20*(8), 490–497. https://doi.org/10.1016/j.tplants.2015.04.012

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., … Merrick, J. M. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science (New York, N.Y.)*, *269*(5223), 496–512. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7542800

Förster, J., Famili, I., Fu, P., Palsson, B. Ø., & Nielsen, J. (2003). Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. *Genome Research*, *13*(2), 244–53. https://doi.org/10.1101/gr.234503

Frayn, K. N., Arner, P., & Yki-Järvinen, H. (2006). Fatty acid metabolism in adipose tissue, muscle and liver in health and disease. *Essays in Biochemistry*, *42*, 89–103. https://doi.org/10.1042/bse0420089

Gautier, L., Cope, L., Bolstad, B. M., & Irizarry, R. A. (2004). affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, *20*(3), 307–315. https://doi.org/10.1093/bioinformatics/btg405

Geeven, G., van Kesteren, R. E., Smit, A. B., & de Gunst, M. C. M. (2012). Identification of context-specific gene regulatory networks with GEMULA--gene expression modeling using LAsso. *Bioinformatics*, *28*(2), 214–221. https://doi.org/10.1093/bioinformatics/btr641

Geigenberger, P., Reimholz, R., Geiger, M., Merlo, L., Canale, V., & Stitt, M. (1997). Regulation of sucrose and starch metabolism in potato tubers in response to short-term water deficit. *Planta*, *201*(4), 502–518. https://doi.org/10.1007/s004250050095

Geigenberger, P., & Stitt, M. (1991). A “futile” cycle of sucrose synthesis and degradation is involved in regulating partitioning between sucrose, starch and respiration in cotyledons of germinating Ricinus communis L. seedlings when phloem transport is inhibited. *Planta*, *185*(1), 81–90. https://doi.org/10.1007/BF00194518

Gomes de Oliveira Dal'Molin, C., Quek, L.-E., Saa, P. A., & Nielsen, L. K. (2015). A multi-tissue genome-scale metabolic modeling framework for the analysis of whole plant systems. *Frontiers in Plant Science*, *6*, 4. https://doi.org/10.3389/fpls.2015.00004

Gotow, K., Taylor, S., & Zeiger, E. (1988). Photosynthetic Carbon Fixation in Guard Cell Protoplasts of Vicia faba L . 1, 700–705.

Goutelle, S., Maurin, M., Rougier, F., Barbaut, X., Bourguignon, L., Ducher, M., & Maire, P. (2008). The Hill equation: a review of its capabilities in pharmacological modelling. *Fundamental & Clinical Pharmacology*, *22*(6), 633–648. https://doi.org/10.1111/j.1472-8206.2008.00633.x

Grafahrend-Belau, E., Junker, A., Eschenröder, A., Müller, J., Schreiber, F., & Junker, B. H. (2013). Multiscale metabolic modeling: dynamic flux balance analysis on a whole-plant scale. *Plant Physiology*, *163*, 637–47. https://doi.org/10.1104/pp.113.224006

Gurobi Optimization, I. (2017). Gurobi Optimizer Reference Manual. Retrieved from http://www.gurobi.com

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., … Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, *8*(8), 1494–1512. https://doi.org/10.1038/nprot.2013.084

Hampp, R., Outlaw, W. H., & Tarczynski, M. C. (1982). Profile of basic carbon pathways in guard cells and other leaf cells of Vicia faba L. *Plant Physiology*, *70*(6), 1582–1585. https://doi.org/10.1104/pp.70.6.1582

Hargreaves, J. A., & ap Rees, T. (1988). Turnover of starch and sucrose in roots of Pisum sativum. *Phytochemistry*, *27*(6), 1627–1629. https://doi.org/10.1016/0031-9422(88)80416-3

Haug, K., Salek, R. M., Conesa, P., Hastings, J., de Matos, P., Rijnbeek, M., … Steinbeck, C. (2013). MetaboLights--an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Research*, *41*(D1), D781–D786. https://doi.org/10.1093/nar/gks1004

Heinrich, R., & Schuster, S. (1996). *The Regulation of Cellular Systems*. Boston, MA: Springer US. https://doi.org/10.1007/978-1-4613-1161-4

Hesterberg, T., Choi, N. H., Meier, L., & Fraley, C. (2008). Least Angle and L1 Penalized Regression: A Review. *Statistics Surveys*, *2*, 61–93. https://doi.org/10.1214/08-SS035

Hetherington, A. M., & Woodward, F. I. (2003). The role of stomata in sensing and driving environmental change. *Nature*, *424*(August), 901–908. https://doi.org/10.1017/CBO9781139165266

Hill, S. A., & ap Rees, T. (1993). Fluxes of carbohydrate metabolism in ripening bananas. *Planta*, *192*(1), 52–60. https://doi.org/10.1007/BF00198692

Hills, A., Chen, Z.-H., Amtmann, A., Blatt, M. R., & Lew, V. L. (2012). OnGuard, a computational platform for quantitative kinetic modeling of guard cell physiology. *Plant Physiology*, *159*(3), 1026–42. https://doi.org/10.1104/pp.112.197244

Holzhütter, H.-G. (2004). The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks. *European Journal of Biochemistry*, *271*(14), 2905–2922. https://doi.org/10.1111/j.1432-1033.2004.04213.x

Horvat, P., Koller, M., & Braunegg, G. (2015). Recent advances in elementary flux modes and yield space analysis as useful tools in metabolic network studies. *World Journal of Microbiology and Biotechnology*, *31*(9), 1315–1328. https://doi.org/10.1007/s11274-015-1887-1

Hubbell, E., Liu, W. M., & Mei, R. (2002). Robust estimators for expression analysis. *Bioinformatics*, *18*(12), 1585–1592. https://doi.org/10.1093/bioinformatics/18.12.1585

Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., … Wang, J. (2003). The

systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, *19*(4), 524–531. https://doi.org/10.1093/bioinformatics/btg015

Huege, J., Goetze, J., Dethloff, F., Junker, B., & Kopka, J. (2014). Plant Chemical Genomics. In G. R. Hicks & S. Robert (Eds.), *Glenn R. Hicks and Stéphanie Robert (eds.), Plant Chemical Genomics: Methods and Protocols, Methods in Molecular Biology, vol. 1056, DOI 10.1007/978-1-62703-592-7* (Vol. 1056, pp. 11–17). Totowa, NJ: Humana Press. https://doi.org/10.1007/978-1-62703-592-7

Human Genome Sequencing Consortium, I. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, *431*(7011), 931–945. https://doi.org/10.1038/nature03001

Hyduke, D. R., Lewis, N. E., & Palsson, B. Ø. (2013). Analysis of omics data with genome-scale models of metabolism. *Molecular bioSystems*, *9*(2), 167–74. https://doi.org/10.1039/c2mb25453k

Hyduke, D., Schellenberger, J., Que, R., Fleming, R., Thiele, I., Orth, J., … Palsson, B. (2011). COBRA Toolbox 2.0. *Protocol Exchange*. https://doi.org/10.1038/protex.2011.234

IBM, I. (n.d.). IBM ILOG CPLEX Solver. Retrieved from http://www.aimms.com/aimms/solvers/cplex/

International Union of Biochemistry and Molecular Biology. Nomenclature Committee., & Webb, E. C. (Edwin C. (1992). *Enzyme nomenclature 1992 : recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes*. Published for the International Union of Biochemistry and Molecular Biology by Academic Press.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., & Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, *4*(2), 249–264. https://doi.org/10.1093/biostatistics/4.2.249

Jensen, P. A., Lutz, K. A., & Papin, J. A. (2011). TIGER: Toolbox for integrating genome-scale metabolic models, expression data, and transcriptional regulatory networks. *BMC Systems Biology*. https://doi.org/10.1186/1752-0509-5-147

Jerby, L., Shlomi, T., & Ruppin, E. (2010). Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. *Molecular Systems Biology*, *6*, 401.

Jiao, J. -a., & Chollet, R. (1991). Posttranslational Regulation of Phosphoenolpyruvate Carboxylase in C4 and Crassulacean Acid Metabolism Plants. *Plant Physiology*, *95*(4), 981–985. https://doi.org/10.1104/pp.95.4.981

Johnson, C. H., Ivanisevic, J., & Siuzdak, G. (2016). Metabolomics: beyond biomarkers and towards mechanisms. *Nature Reviews Molecular Cell Biology*, *17*(7), 451–459. https://doi.org/10.1038/nrm.2016.25

Johnson, K. A., Goody, R. S., Johnson, K. A., & Goody, R. S. (2011). The Original Michaelis Constant: Translation of the 1913 Michaelis–Menten Paper. *Biochemistry*, *50*(39), 8264–8269. https://doi.org/10.1021/bi201284u

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, *44*(D1), D457–D462. https://doi.org/10.1093/nar/gkv1070

Karr, J. R., Sanghvi, J. C., Macklin, D. N., Gutschow, M. V, Jacobs, J. M., Bolival, B., … al., et. (2012). A whole-cell computational model predicts phenotype from genotype. *Cell*, *150*(2), 389–401. https://doi.org/10.1016/j.cell.2012.05.044

Kelk, S. M., Olivier, B. G., Stougie, L., & Bruggeman, F. J. (2012). Optimal flux spaces of genome-scale stoichiometric models are determined by a few subnetworks. *Scientific Reports*, *2*, 580. https://doi.org/10.1038/srep00580

Kelly, G., Moshelion, M., David-Schwartz, R., Halperin, O., Wallach, R., Attia, Z., … Granot, D. (2013).

Hexokinase mediates stomatal closure. *Plant Journal*, *75*(6), 977–988. https://doi.org/10.1111/tpj.12258

Kim, T. Y., Sohn, S. B., Kim, Y. Bin, & Kim, W. J. (2012). Recent advances in reconstruction and applications of genome-scale metabolic models. *Current Opinion in Biotechnology*, *23*(4), 617–623. https://doi.org/10.1016/j.copbio.2011.10.007

King, Z. A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J. A., … Lewis, N. E. (2016). BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Research*, *44*(D1), D515–D522. https://doi.org/10.1093/nar/gkv1049

Koch, L. (2016). Complex disease: A global view of regulatory networks. *Nature Reviews Genetics*, *17*(5), 252–252. https://doi.org/10.1038/nrg.2016.36

Kopka, J., Schauer, N., Krueger, S., Birkemeyer, C., Usadel, B., Bergmüller, E., … Steinhauser, D. (2005). GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics (Oxford, England)*, *21*(8), 1635–8. https://doi.org/10.1093/bioinformatics/bti236

Krall, L., Huege, J., Catchpole, G., Steinhauser, D., & Willmitzer, L. (2009). Assessment of sampling strategies for gas chromatography-mass spectrometry (GC-MS) based metabolomics of cyanobacteria. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences*, *877*(27), 2952–2960. https://doi.org/10.1016/j.jchromb.2009.07.006

Kruger, N. J., Masakapalli, S. K., & Ratcliffe, R. G. (2012). Strategies for investigating the plant metabolic network with steady-state metabolic flux analysis: Lessons from an Arabidopsis cell culture and other systems. *Journal of Experimental Botany*. https://doi.org/10.1093/jxb/err382

Krupp, M., Marquardt, J. U., Sahin, U., Galle, P. R., Castle, J., & Teufel, A. (2012). RNA-Seq Atlas-a reference database for gene expression profiling in normal tissue by next-generation sequencing. *Bioinformatics*, *28*, 1184–1185. https://doi.org/10.1093/bioinformatics/bts084

Lash, L. H. (2005). Role of glutathione transport processes in kidney function. *Toxicology and Applied Pharmacology*. https://doi.org/10.1016/j.taap.2004.10.004

Lawrence, K. D., & Shier, D. R. . (2010). A comparison of least squares and least absolute deviation regression models for estimating Weibull parameters. *Communications in Statistics - Simulation and Computation*. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/03610919808813515a

Lawson, T. (2009). Guard cell photosynthesis and stomatal function. *The New Phytologist*, *181*(1), 13–34. https://doi.org/10.1111/j.1469-8137.2008.02685.x

Lawson, T., & Blatt, M. R. (2014). Stomatal size, speed, and responsiveness impact on photosynthesis and water use efficiency. *Plant Physiol.*, *164*(4), 1556–1570. https://doi.org/10.1104/pp.114.237107

Lawson, T., Oxborough, K., Morison, J. I. L., & Baker, N. R. (2002). Responses of photosynthetic electron transport in stomatal guard cells and mesophyll cells in intact leaves to light, CO2, and humidity. *Plant Physiology*, *128*(1), 52–62. https://doi.org/10.1104/pp.010317

Lawson, T., Oxborough, K., Morison, J. I. L., & Baker, N. R. (2003). The responses of guard and mesophyll cell photosynthesis to Co2, O2, light, and water stress in a range of species are similar. *Journal of Experimental Botany*, *54*(388), 1743–1752. https://doi.org/10.1093/jxb/erg186

Lawson, T., Simkin, A. J., Kelly, G., & Granot, D. (2014). Mesophyll photosynthesis and guard cell metabolism impacts on stomatal behaviour. *New Phytologist*, *203*(4), 1064–1081. https://doi.org/10.1111/nph.12945

Lee, D., Smallbone, K., Dunn, W. B., Murabito, E., Winder, C. L., Kell, D. B., … Swainston, N. (2012). Improving metabolic flux predictions using absolute gene expression data. *BMC Systems Biology*, *6*(1), 73. https://doi.org/10.1186/1752-0509-6-73

Lee, S., Phalakornkule, C., Domach, M. M., & Grossmann, I. E. (2000). Recursive MILP model for

finding all the alternate optima in LP models for metabolic networks. *Computers & Chemical Engineering*, *24*(2–7), 711–716. https://doi.org/10.1016/S0098-1354(00)00323-9

Leonhardt, N., Kwak, J. M., Robert, N., Waner, D., Leonhardt, G., & Schroeder, J. I. (2004). Microarray expression analyses of Arabidopsis guard cells and isolation of a recessive abscisic acid hypersensitive protein phosphatase 2C mutant. *The Plant Cell*, *16*(3), 596–615. https://doi.org/10.1105/tpc.019000

Leskovac, V. (2003). *Comprehensive enzyme kinetics*. Kluwer Academic/Plenum Pub.

Lewis, N. E., Cho, B.-K., Knight, E. M., & Palsson, B. O. (2009). Gene expression profiling and the use of genome-scale in silico models of Escherichia coli for analysis: providing context for content. *Journal of Bacteriology*, *191*(11), 3437–44. https://doi.org/10.1128/JB.00034-09

Lewis, N. E., Nagarajan, H., & Palsson, B. O. (2012). Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. *Nature Reviews Microbiology*. https://doi.org/10.1038/nrmicro2737

Li, S., Assmann, S. M., & Albert, R. (2006). Predicting essential components of signal transduction networks: A dynamic model of guard cell abscisic acid signaling. *PLoS Biology*, *4*(10), 1732–1748. https://doi.org/10.1371/journal.pbio.0040312

Lisec, J., Schauer, N., Kopka, J., Willmitzer, L., & Fernie, A. R. (2006). Gas chromatography mass spectrometry–based metabolite profiling in plants. *Nature Protocols*, *1*(1), 387–396. https://doi.org/10.1038/nprot.2006.59

Llaneras, F., & Pic??, J. (2010). Which metabolic pathways generate and characterize the flux space? A comparison among elementary modes, extreme pathways and minimal generators. *Journal of Biomedicine and Biotechnology*. https://doi.org/10.1155/2010/753904

Lovric, M. (Ed.). (2011). *International Encyclopedia of Statistical Science*. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-04898-2

Lugassi, N., Kelly, G., Fidel, L., Yaniv, Y., Attia, Z., Levi, A., … Granot, D. (2015). Expression of Arabidopsis Hexokinase in Citrus Guard Cells Controls Stomatal Aperture and Reduces Transpiration. *Frontiers in Plant Science*, *6*(December), 1114. https://doi.org/10.3389/fpls.2015.01114

Lunn, J. E. (2008). Sucrose Metabolism. In *Encyclopedia of Life Sciences* (Vol. 110, p. 43). Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470015902.a0021259

Ma, F., Jazmin, L. J., Young, J. D., & Allen, D. K. (2014). Isotopically nonstationary 13C flux analysis of changes in Arabidopsis thaliana leaf metabolism due to high light acclimation. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(47), 16967–72. https://doi.org/10.1073/pnas.1319485111

Ma, S., Minch, K. J., Rustad, T. R., Hobbs, S., Zhou, S.-L., Sherman, D. R., & Price, N. D. (2015). Integrated Modeling of Gene Regulatory and Metabolic Networks in Mycobacterium tuberculosis. *PLOS Computational Biology*, *11*(11), e1004543. https://doi.org/10.1371/journal.pcbi.1004543

Machado, D., & Herrgård, M. (2014). Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. *PLoS Comput Biol*, *10*(4), e1003580. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24762745

Machado, D., Herrgård, M. J., & Rocha, I. (2016). Stoichiometric Representation of Gene-Protein-Reaction Associations Leverages Constraint-Based Analysis from Reaction to Gene-Level Phenotype Prediction. *PLoS Computational Biology*, *12*(10), e1005140. https://doi.org/10.1371/journal.pcbi.1005140

Magnúsdóttir, S., Heinken, A., Kutt, L., Ravcheev, D. A., Bauer, E., Noronha, A., … Thiele, I. (2017). Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nature Biotechnology*, *35*(1), 81–89. https://doi.org/10.1038/nbt.3703

Mahadevan, R., Edwards, J. S., & Doyle, F. J. (2002). Dynamic Flux Balance Analysis of Diauxic Growth in Escherichia coli. *Biophysical Journal*, *83*(3), 1331–1340. https://doi.org/10.1016/S0006-3495(02)73903-9

Mahadevan, R., & Schilling, C. H. (2003). The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metabolic Engineering*, *5*(4), 264–276.

Makhorin, A. (2012). GNU Linear Programming Kit (GLPK). Retrieved from https://www.gnu.org/software/glpk/

Mallmann, J., Heckmann, D., Bräutigam, A., Lercher, M. J., Weber, A. P. M., Westhoff, P., & Gowik, U. (2014). The role of photorespiration during the evolution of C4 photosynthesis in the genus Flaveria. *eLife*, *3*, e02478. https://doi.org/10.7554/eLife.02478

Marmiesse, L., Peyraud, R., & Cottret, L. (2015). FlexFlux: combining metabolic flux and regulatory network analyses. *BMC Systems Biology*, *9*(1), 93. https://doi.org/10.1186/s12918-015-0238-z

Marx, V. (2014). Proteomics: An atlas of expression. *Nature*, *509*(7502), 645–9. https://doi.org/10.1038/509645a

McCall, M. N., Jaffee, H. A., Zelisko, S. J., Sinha, N., Hooiveld, G., Irizarry, R. A., & Zilliox, M. J. (2014). The Gene Expression Barcode 3.0: Improved data processing and mining tools. *Nucleic Acids Research*, *42*. https://doi.org/10.1093/nar/gkt1204

McDonald, G. C. (2009). Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, *1*, 93–100. https://doi.org/10.1002/wics.14

Medeiros, D. B., Daloso, D. M., Fernie, A. R., Nikoloski, Z., & Ara??jo, W. L. (2015). Utilizing systems biology to unravel stomatal function and the hierarchies underpinning its control. *Plant, Cell and Environment*, *38*(8), 1457–1470. https://doi.org/10.1111/pce.12517

Melzer, E., & O'leary, M. H. (1987). Anapleurotic CO(2) Fixation by Phosphoenolpyruvate Carboxylase in C(3) Plants. *Plant Physiology*, *84*(1), 58–60. https://doi.org/10.1104/pp.84.1.58

Metallo, C. M., & Vander Heiden, M. G. (2013). Understanding metabolic regulation and its influence on cell physiology. *Molecular Cell*, *49*(3), 388–98. https://doi.org/10.1016/j.molcel.2013.01.018

Minguet-Parramona, C., Wang, Y., Hills, A., Vialet-Chabrand, S., Griffiths, H., Rogers, S., … Blatt, M. R. (2016). An Optimal Frequency in Ca2+ Oscillations for Stomatal Closure Is an Emergent Property of Ion Transport in Guard Cells. *Plant Physiol.*, *170*(1), 33–42. https://doi.org/10.1104/pp.15.01607

Mintz-Oron, S., Meir, S., Malitsky, S., Ruppin, E., Aharoni, A., & Shlomi, T. (2012). Reconstruction of Arabidopsis metabolic network models accounting for subcellular compartmentalization and tissue-specificity. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1100358109

Misra, B. B., Acharya, B. R., Granot, D., Assmann, S. M., & Chen, S. (2015). The guard cell metabolome: functions in stomatal movement and global food security. *Frontiers in Plant Science*, *6*, 334. https://doi.org/10.3389/fpls.2015.00334

Misra, B. B., De Armas, E., Tong, Z., & Chen, S. (2015). Metabolomic Responses of Guard Cells and Mesophyll Cells to Bicarbonate. *PLoS ONE*, *10*(12). https://doi.org/10.1371/journal.pone.0144206

Mitra, V., & Metcalf, J. (2009). Metabolic functions of the liver. *Anaesthesia and Intensive Care Medicine*. https://doi.org/10.1016/j.mpaic.2009.03.011

Moretti, S., Martin, O., Van Du Tran, T., Bridge, A., Morgat, A., & Pagni, M. (2016). MetaNetX/MNXref – reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. *Nucleic Acids Research*, *44*(D1), D523–D526. https://doi.org/10.1093/nar/gkv1117

Moxley, J. F., Jewett, M. C., Antoniewicz, M. R., Villas-Boas, S. G., Alper, H., Wheeler, R. T., … Stephanopoulos, G. (2009). Linking high-resolution metabolic flux phenotypes and transcriptional regulation in yeast modulated by the global regulator Gcn4p. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 6477–6482. https://doi.org/10.1073/pnas.0811091106

Moyano, T. C., Vidal, E. A., Contreras-López, O., & Gutiérrez, R. A. (2015). Constructing simple biological networks for understanding complex high-throughput data in plants. *Methods in Molecular Biology (Clifton, N.J.)*, *1284*, 503–26. https://doi.org/10.1007/978-1-4939-2444-8_25

Müller, A. C., & Bockmayr, A. (2014). Flux modules in metabolic networks. *Journal of Mathematical Biology*, *69*(5), 1151–79. https://doi.org/10.1007/s00285-013-0731-1

Nargund, S., Misra, A., Zhang, X., Coleman, G. D., & Sriram, G. (2014). Flux and reflux: metabolite reflux in plant suspension cells and its implications for isotope-assisted metabolic flux analysis. *Molecular bioSystems*, *10*(6), 1496–508. https://doi.org/10.1039/c3mb70348g

Ni, D. A. (2012). Role of vacuolar invertase in regulating Arabidopsis stomatal opening. *Acta Physiologiae Plantarum*, *34*(6), 2449–2452. https://doi.org/10.1007/s11738-012-1036-5

Niedenführ, S., Wiechert, W., & Nöh, K. (2015). How to measure metabolic fluxes: a taxonomic guide for 13C fluxomics. *Current Opinion in Biotechnology*, *34*, 82–90. https://doi.org/10.1016/j.copbio.2014.12.003

Nikoloski, Z., Perez-Storey, R., & Sweetlove, L. J. (2015). Inference and Prediction of Metabolic Network Fluxes. *Plant Physiol*, *169*(3), 1443–1455. https://doi.org/10.1104/pp.15.01082

Nunes-Nesi, A., Carrari, F., Gibon, Y., Sulpice, R., Lytovchenko, A., Fisahn, J., … Fernie, A. R. (2007). Deficiency of mitochondrial fumarase activity in tomato plants impairs photosynthesis via an effect on stomatal function. *Plant Journal*, *50*(6), 1093–1106. https://doi.org/10.1111/j.1365-313X.2007.03115.x

Nuzhdin, S. V, Friesen, M. L., & McIntyre, L. M. (2012). Genotype-phenotype mapping in a post-GWAS world. *Trends in Genetics : TIG*, *28*(9), 421–6. https://doi.org/10.1016/j.tig.2012.06.003

Oberhardt, M. A., Palsson, B. Ø., & Papin, J. A. (2009). Applications of genome-scale metabolic reconstructions. *Molecular Systems Biology*, *5*, 320. https://doi.org/10.1038/msb.2009.77

Orth, J. D., Thiele, I., & Palsson, B. Ø. (2010). What is flux balance analysis? *Nature Biotechnology*, *28*, 245–248. https://doi.org/10.1038/nbt.1614

Osmond, C. B. (1978). Crassulacean Acid Metabolism : a Curiosity. *Annual Review of Plant Physiology*, *29*, 379–414. https://doi.org/10.1146/annurev.pp.29.060178.002115

Outlaw, W. H. J. (2003). Critical Reviews in Plant Sciences Integration of Cellular and Physiological Functions of Guard Cells Integration of Cellular and Physiological Functions of Guard Cells. *Critical Reviews in Plant Sciences*, *22*(6), 503–5229. https://doi.org/10.1080/07352680390253511

Outlaw, W. H., & Kennedy, J. (1978). Enzymic and substrate basis for the anaplerotic step in guard cells. *Plant Physiology*, *62*(4), 648–652.

Pacheco, M. P., John, E., Kaoma, T., Heinäniemi, M., Nicot, N., Vallar, L., … Sauter, T. (2015). Integrated metabolic modelling reveals cell-type specific epigenetic control points of the macrophage metabolic network. *BMC Genomics*, *16*(1), 809. https://doi.org/10.1186/s12864-015-1984-4

Pacheco, M. P., Pfau, T., & Sauter, T. (2016). Benchmarking procedures for high-throughput context specific reconstruction algorithms. *Frontiers in Physiology*, *6*(JAN). https://doi.org/10.3389/fphys.2015.00410

Palsson, B. (2006). *Systems biology : properties of reconstructed networks*. Cambridge University Press. Retrieved from http://www.cambridge.org/gb/academic/subjects/life-sciences/genomics-

bioinformatics-and-systems-biology/systems-biology-properties-reconstructed-networks?format=HB&isbn=9780521859035#UTRKdba23TSJfsbe.97

Pan, W., Yuan, Y., & Stan, G. (2012). Reconstruction of arbitrary biochemical reaction networks: a compressive sensing approach. *CoRR*, *abs/1205.1*, 1–15. Retrieved from http://arxiv.org/abs/1205.1720

Pandey, S., Wang, R.-S., Wilson, L., Li, S., Zhao, Z., Gookin, T. E., … Albert, R. (2010). Boolean modeling of transcriptome data reveals novel modes of heterotrimeric G-protein action. *Molecular Systems Biology*, *6*, 372. https://doi.org/10.1038/msb.2010.28

Parvanthi, K., & Raghavendra, A. S. (1997). Both rubisco and. *PLant Science*, *124*, 153–157.

Penfield, S., Clements, S., Bailey, K. J., Gilday, A. D., Leegood, R. C., Gray, J. E., & Graham, I. A. (2012). Expression and manipulation of PHOSPHOENOLPYRUVATE CARBOXYKINASE 1 identifies a role for malate metabolism in stomatal closure. *Plant Journal*, *69*(4), 679–688. https://doi.org/10.1111/j.1365-313X.2011.04822.x

Petryszak, R., Burdett, T., Fiorelli, B., Fonseca, N. A., Gonzalez-Porta, M., Hastings, E., … Brazma, A. (2014). Expression Atlas update--a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Research*, *42*(Database issue), D926-32. https://doi.org/10.1093/nar/gkt1270

Pigliucci, M. (2010). Genotype–phenotype mapping and the end of the "genes as blueprint" metaphor. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *365*(1540).

Placzek, S., Schomburg, I., Chang, A., Jeske, L., Ulbrich, M., Tillack, J., & Schomburg, D. (2017). BRENDA in 2017: new perspectives and new tools in BRENDA. *Nucleic Acids Research*, *45*(D1), D380–D388. https://doi.org/10.1093/nar/gkw952

Poolman, M. G., Miguet, L., Sweetlove, L. J., & Fell, D. A. (2009). A genome-scale metabolic model of Arabidopsis and some of its properties. *Plant Physiology*, *151*, 1570–1581. https://doi.org/10.1104/pp.109.141267

Pornputtapong, N., Nookaew, I., & Nielsen, J. (2015). Human metabolic atlas: an online resource for human metabolism. *Database : The Journal of Biological Databases and Curation*, *2015*, bav068. https://doi.org/10.1093/database/bav068

Ravikirthi, P., Suthers, P. F., & Maranas, C. D. (2011). Construction of an E. Coli genome-scale atom mapping model for MFA calculations. *Biotechnology and Bioengineering*, *108*, 1372–1382. https://doi.org/10.1002/bit.23070

Recht, L., Töpfer, N., Batushansky, A., Sikron, N., Gibon, Y., Fait, A., … Zarka, A. (2014). Metabolite Profiling and Integrative Modeling Reveal Metabolic Constraints for Carbon Partitioning under Nitrogen-Starvation in the Green Alga Haematococcus pluvialis. *The Journal of Biological Chemistry*, *289*, 30387–30403. https://doi.org/10.1074/jbc.M114.555144

Reckmann, U., Scheibe, R., & Raschke, K. (1990). Rubisco activity in guard cells compared with the solute requirement for stomatal opening. *Plant Physiology*, *92*(1), 246–53. https://doi.org/10.1104/pp.92.1.246

Reed, J. L., & Palsson, B. Ø. (2004). Genome-scale in silico models of E. coli have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states. *Genome Research*, *14*(9), 1797–805. https://doi.org/10.1101/gr.2546004

Reimers, A.-M., & Reimers, A. C. (2016). The steady-state assumption in oscillating and growing systems. *Journal of Theoretical Biology*, *406*, 176–186. https://doi.org/10.1016/j.jtbi.2016.06.031

Righetti, P., Campostrini, N., Pascali, J., Hamdan, M., & Astner, H. (2004). Quantitative proteomics: a review of different methodologies. *European Journal of Mass Spectrometry*, *10*(1), 335. https://doi.org/10.1255/ejms.600

Robaina Estévez, S., & Nikoloski, Z. (2014). Generalized framework for context-specific metabolic model extraction methods. *Frontiers in Plant Science*, *5*(September), 491. https://doi.org/10.3389/fpls.2014.00491

Robaina Estévez, S., & Nikoloski, Z. (2015). Context-Specific Metabolic Model Extraction Based on Regularized Least Squares Optimization. *PloS One*, *10*(7), e0131875. https://doi.org/10.1371/journal.pone.0131875

Robaina Estévez, S., & Nikoloski, Z. (2017). On the effects of alternative optima in context-specific metabolic model predictions. *PLoS Computational Biology*.

Rochfort, S. (2005). Metabolomics Reviewed: A New "Omics" Platform Technology for Systems Biology and Implications for Natural Products Research. *Journal of Natural Products*, *68*(12), 1813–1820. https://doi.org/10.1021/np050255w

Rockafellar, R. T. (n.d.). *Convex analysis*.

Roessner, U., Luedemann, a, Brust, D., Fiehn, O., Linke, T., Willmitzer, L., & Fernie, a. (2001). Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *The Plant Cell*, *13*(1), 11–29. https://doi.org/10.1105/tpc.13.1.11

Rossell, S., Huynen, M. A., & Notebaart, R. A. (2013). Inferring Metabolic States in Uncharacterized Environments Using Gene-Expression Measurements. *PLoS Computational Biology*, *9*(3). https://doi.org/10.1371/journal.pcbi.1002988

Rossell, S., van der Weijden, C. C., Lindenbergh, A., van Tuijl, A., Francke, C., Bakker, B. M., & Westerhoff, H. V. (2006). Unraveling the complexity of flux regulation: a new method demonstrated for nutrient starvation in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America*, *103*, 2166–2171. https://doi.org/10.1073/pnas.0509831103

Santelia, D., & Lawson, T. (2016). Rethinking guard cell metabolism. *Plant Physiology*. https://doi.org/10.1104/pp.16.00767

Savageau, M. A. (1969). Biochemical systems analysis. *Journal of Theoretical Biology*, *25*(3), 370–379. https://doi.org/10.1016/S0022-5193(69)80027-5

Savinell, J. M., & Palsson, B. O. (1992). Network analysis of intermediary metabolism using linear optimization. I. Development of mathematical formalism. *Journal of Theoretical Biology*, *154*(4), 421–454. https://doi.org/10.1016/S0022-5193(05)80161-4

Schellenberger, J., & Palsson, B. Ø. (2009). Use of randomized sampling for analysis of metabolic networks. *The Journal of Biological Chemistry*, *284*(9), 5457–5461. https://doi.org/10.1074/jbc.R800048200

Schellenberger, J., Park, J. O., Conrad, T. M., & Palsson, B. Ø. (2010). BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics*, *11*, 213. https://doi.org/10.1186/1471-2105-11-213

Schmidt, B. J., Ebrahim, A., Metz, T. O., Adkins, J. N., Palsson, B., & Hyduke, D. R. (2013). GIM3E: Condition-specific models of cellular metabolism developed from metabolomics and expression data. *Bioinformatics*, *29*(22), 2900–2908.

Schultz, A., & Qutub, A. A. (2016). Reconstruction of Tissue-Specific Metabolic Networks Using CORDA. *PLoS Computational Biology*, *12*(3), e1004808. https://doi.org/10.1371/journal.pcbi.1004808

Schwender, J., Ohlrogge, J., & Shachar-Hill, Y. (2004). Understanding flux in plant metabolic networks. *Current Opinion in Plant Biology*, *7*(3), 309–317. https://doi.org/10.1016/j.pbi.2004.03.016

Seaver, S. M. D., Gerdes, S., Frelin, O., Lerma-Ortiz, C., Bradbury, L. M. T., Zallot, R., … Henry, C. S. (2014). High-throughput comparison, functional annotation, and metabolic modeling of plant

genomes using the PlantSEED resource. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(26), 9645–50. https://doi.org/10.1073/pnas.1401329111

Segrè, D., Vitkup, D., & Church, G. M. (2002). Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(23), 15112–7. https://doi.org/10.1073/pnas.232349399

Sharkey, T. D. (1988). Estimating the rate of photorespiration in leaves. *Physiologia Plantarum*, *73*(1), 147–152. https://doi.org/10.1111/j.1399-3054.1988.tb09205.x

Sheikh, K., Förster, J., & Nielsen, L. K. (2008). Modeling Hybridoma Cell Metabolism Using a Generic Genome-Scale Metabolic Model of Mus musculus. *Biotechnology Progress*, *21*(1), 112–121. https://doi.org/10.1021/bp0498138

Shimazaki, K., Terada, J., Tanaka, K., & Kondo, N. (1989). Calvin-Benson Cycle Enzymes in Guard-Cell Protoplasts from Vicia faba L. *Plant Physiology*, *90*(3), 1057–1064.

Shimazaki, K., & Zeiger, E. (1985). Cyclic and Noncyclic Photophosphorylation in Isolated Guard Cell Chloroplasts from Vicia.faba L. *Plant Physiol.*, *78*, 211–214.

Shlomi, T., Cabili, M. N., Herrgård, M. J., Palsson, B. Ø., & Ruppin, E. (2008). Network-based prediction of human tissue-specific metabolism. *Nature Biotechnology*, *26*(9), 1003–1010. https://doi.org/10.1038/nbt.1487

Stein, L. (2001). Genome annotation: from sequence to biology. *Nature Reviews Genetics*, *2*(7), 493–503. https://doi.org/10.1038/35080529

Suetsugu, N., Takami, T., Ebisu, Y., Watanabe, H., Iiboshi, C., Doi, M., & Shimazaki, K. I. (2014). Guard cell chloroplasts are essential for blue light-dependent stomatal opening in arabidopsis. *PLoS ONE*, *9*(9). https://doi.org/10.1371/journal.pone.0108374

Sun, Z., Jin, X., Albert, R., & Assmann, S. M. (2014). Multi-level Modeling of Light-Induced Stomatal Opening Offers New Insights into Its Regulation by Drought. *PLoS Computational Biology*, *10*(11), e1003930. https://doi.org/10.1371/journal.pcbi.1003930

Szecowka, M., Heise, R., Tohge, T., Nunes-Nesi, A., Vosloh, D., Huege, J., … Arrivault, S. (2013). Metabolic fluxes in an illuminated Arabidopsis rosette. *The Plant Cell*, *25*(2), 694–714. https://doi.org/10.1105/tpc.112.106989

Talbott, L. D., & Zeiger, E. (1993). Sugar and Organic Acid Accumulation in Guard Cells of Vicia faba in Response to Red and Blue Light. *Plant Physiology*, *102*(90), 1163–1169. https://doi.org/10.1104/pp.102.4.1163

Talbott, L., & Zeiger, E. (1998). The role of sucrose in guard cell osmoregulation. *Journal of Experimental Botany*, *49*(90001), 329–337. https://doi.org/10.1093/jexbot/49.suppl_1.329

Terzer, M. (2009). Large scale methods to enumerate extreme rays and elementary modes. https://doi.org/10.3929/ETHZ-A-005945733

Thiele, I., & Palsson, B. Ø. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols*, *5*(1), 93–121. https://doi.org/10.1038/nprot.2009.203

Thiele, I., Swainston, N., Fleming, R. M. T., Hoppe, A., Sahoo, S., Aurich, M. K., … Palsson, B. Ø. (2013). A community-driven global reconstruction of human metabolism. *Nature Biotechnology*, *31*, 419–25. https://doi.org/10.1038/nbt.2488

Tibshirani, R. (1994). Regression Selection and Shrinkage via the Lasso. *Journal of the Royal Statistical Society B*, *58*, 267–288. https://doi.org/10.2307/2346178

Tirthankar Sengupta. Shivi Jain. Mani Bhushan. (2013). A Compressed Sensing Based Basis-pursuit Formulation of the Room Algorithm. *Preprints of the 12th IFAC Symposium on Computer Applications in Biothecnology. The International Federation of Automatic Control.*, 16–18.

Töpfer, N., Kleessen, S., & Nikoloski, Z. (2015). Integration of metabolomics data profiles into metabolic networks. *Frontiers in Plant Science*, *6*(49). https://doi.org/10.3389/fpls.2015.00049

Trethewey, R. N., Geigenberger, P., Riedel, K., Hajirezaei, M. R., Sonnewald, U., Stitt, M., … Willmitzer, L. (1998). Combined expression of glucokinase and invertase in potato tubers leads to a dramatic reduction in starch accumulation and a stimulation of glycolysis. *Plant Journal*, *15*(1), 109–118. https://doi.org/10.1046/j.1365-313X.1998.00190.x

Uhlén, M., Björling, E., Agaton, C., Szigyarto, C. A.-K., Amini, B., Andersen, E., … Pontén, F. (2005). A human protein atlas for normal and cancer tissues based on antibody proteomics. *Molecular & Cellular Proteomics : MCP*, *4*(12), 1920–32. https://doi.org/10.1074/mcp.M500279-MCP200

Uhlén, M. et al. (2015). Tissue-based map of the human proteome. *Science*, *347*(6220). https://doi.org/10.1126/science.1260419

Uhlen, M., Hallstro m, B. M., Lindskog, C., Mardinoglu, A., Ponten, F., & Nielsen, J. (2016). Transcriptomics resources of human tissues and organs. *Molecular Systems Biology*, *12*(4), 862–862. https://doi.org/10.15252/msb.20155865

Varma, A., & Palsson, B. O. (1994). Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. *Applied and Environmental Microbiology*, *60*(10), 3724–31. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7986045

Vavasseur, A., & Raghavendra, A. S. (2005). Guard cell metabolism and CO_2 sensing. *New Phytologist*, *165*(3), 665–682. https://doi.org/10.1111/j.1469-8137.2004.01276.x

Vidaurre, D., Bielza, C., & Larrañaga, P. (2013). A Survey of L 1 Regression. *International Statistical Review*, n/a-n/a. https://doi.org/10.1111/insr.12023

Vinaixa, M., Rodríguez, M. A., Aivio, S., Capellades, J., Gómez, J., Canyellas, N., … Yanes, O. (2017). Positional Enrichment by Proton Analysis (PEPA): A One-Dimensional ^{1}H-NMR Approach for ^{13}C Stable Isotope Tracer Studies in Metabolomics. *Angewandte Chemie International Edition*. https://doi.org/10.1002/anie.201611347

Vlassis, N., Pacheco, M. P., & Sauter, T. (2014). Fast Reconstruction of Compact Context-Specific Metabolic Network Models. *PLoS Computational Biology*, *10*(1).

Vlassis, N., Pires Pacheco, M., & Sauter, T. (2014). FastCORE MATLAB implementation. Retrieved from http://wwwen.uni.lu/recherche/fstc/life_sciences_research_unit/research_areas/systems_biology/software

Voet, D., & Voet, J. G. (2011). *Biochemistry*. Wiley.

Voit, E. O., Martens, H. A., & Omholt, S. W. (2015). 150 Years of the Mass Action Law. *PLoS Computational Biology*, *11*(1), e1004012. https://doi.org/10.1371/journal.pcbi.1004012

Wang, H., Li, G., & Jiang, G. (2007). Robust Regression Shrinkage and Consistent Variable Selection Through the LAD-Lasso. *Journal of Business & Economic Statistics*. https://doi.org/10.1198/073500106000000251

Wang, R.-S., Pandey, S., Li, S., Gookin, T. E., Zhao, Z., Albert, R., & Assmann, S. M. (2011). Common and unique elements of the ABA-regulated transcriptome of Arabidopsis guard cells. *BMC Genomics*, *12*(1), 216. https://doi.org/10.1186/1471-2164-12-216

Wang, Y., & Blatt, M. R. (2011). Anion channel sensitivity to cytosolic organic acids implicates a central role for oxaloacetate in integrating ion flux with metabolism in stomatal guard cells. *The Biochemical Journal*, *439*(1), 161–70. https://doi.org/10.1042/BJ20110845

Wang, Y., Eddy, J. a, & Price, N. D. (2012). Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. *BMC Systems Biology*, *6*, 153. https://doi.org/10.1186/1752-0509-

6-153

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57–63. https://doi.org/10.1038/nrg2484

Wendler, R., Veith, R., Dancer, J., Stitt, M., & Komor, E. (1991). Sucrose storage in cell suspension cultures of Saccharum sp. (sugarcane) is regulated by a cycle of synthesis and degradation. *Planta*, *183*(1), 31–39. https://doi.org/10.1007/BF00197564

Wiback, S. J., & Palsson, B. O. (2002). Extreme Pathway Analysis of Human Red Blood Cell Metabolism. *Biophysical Journal*, *83*(2), 808–818. https://doi.org/10.1016/S0006-3495(02)75210-7

Willmer, C. M., & Dittrich, P. (1974). Carbon dioxide fixation by epidermal and mesophyll tissues of Tulipa and Commelina. *Planta*, *117*(2), 123–132. https://doi.org/10.1007/BF00390794

Willmer, C. M., & Fricker, M. (1996). *Stomata*. London, UK: Chapman & Hall.

Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y., … Scalbert, A. (2013). HMDB 3.0-The Human Metabolome Database in 2013. *Nucleic Acids Research*, *41*(D1). https://doi.org/10.1093/nar/gks1065

Wu, F.-H., Shen, S.-C., Lee, L.-Y., Lee, S.-H., Chan, M.-T., & Lin, C.-S. (2009). Tape-Arabidopsis Sandwich - a simpler Arabidopsis protoplast isolation method. *Plant Methods*, *5*, 16. https://doi.org/10.1186/1746-4811-5-16

Yang, Y., Costa, A., Leonhardt, N., Siegel, R. S., & Schroeder, J. I. (2008). Isolation of a strong Arabidopsis guard cell promoter and its potential as a research tool. *Plant Methods*, *4*(1), 6. https://doi.org/10.1186/1746-4811-4-6

Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D., … Flicek, P. (2016). Ensembl 2016. *Nucleic Acids Research*, *44*(D1), D710–D716. https://doi.org/10.1093/nar/gkv1157

Young, J. D., Shastri, A. A., Stephanopoulos, G., & Morgan, J. A. (2011). Mapping photoautotrophic metabolism with isotopically nonstationary (13)C flux analysis. *Metabolic Engineering*, *13*(6), 656–65.

Zanghellini, J., Ruckerbauer, D. E., Hanscho, M., & Jungreuthmayer, C. (2013). Elementary flux modes in a nutshell: Properties, calculation and applications. *Biotechnology Journal*, *8*(9), 1009–1016. https://doi.org/10.1002/biot.201200269

Zeiger, E., Talbott, L. D., Frechilla, S., Srivastava, A., & Zhu, J. (2002). The guard cell chloroplast: A perspective for the twenty-first century. *New Phytologist*, *153*(3), 415–424. https://doi.org/10.1046/j.1469-8137.2002.00328.x

Zhang, C., & Hua, Q. (2015). Applications of Genome-Scale Metabolic Models in Biotechnology and Systems Medicine. *Frontiers in Physiology*, *6*, 413. https://doi.org/10.3389/fphys.2015.00413

Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., & Liu, X. (2014). Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. *PLoS ONE*, *9*(1), e78644. https://doi.org/10.1371/journal.pone.0078644

Zhu, J., Talbott, L. D., Jin, X., & Zeiger, E. (1998). The stomatal response to CO2 is linked to changes in guard cell zeaxanthin. *Plant, Cell and Environment*, *21*(8), 813–820. https://doi.org/10.1046/j.1365-3040.1998.00323.x

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *67*, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Zur, H., Ruppin, E., & Shlomi, T. (2010). iMAT: An integrative metabolic analysis tool. *Bioinformatics*, *26*, 3140–3142. https://doi.org/10.1093/bioinformatics/btq602