Advancing elementary flux mode analysis for large-scale metabolic flux networks under steady state

Justin Gareth Chitpin

Thesis submitted to the University of Ottawa in partial fulfillment of the requirements for the PhD degree in Biochemistry

Department of Biochemistry, Microbiology, and Immunology
Faculty of Medicine
University of Ottawa

Abstract

Cellular metabolism is a dynamic process regulating the production and consumption of metabolites. These reaction rates, or metabolic fluxes, are regulated at multiple cellular levels by genes, transcripts, proteins, and metabolites. Advancements in technological and computational methods are leading to increasingly comprehensive estimates of steady state fluxes. In this thesis, I focus on advancing methods to analyse steady state flux data using minimal, steady state pathways known as elementary flux modes (EFMs). Since an arbitrary set of steady state fluxes can be reconstructed by a positive, linear combination of EFMs, these pathways can be viewed as functional units of steady state flow within and through a metabolic flux network. Working with EFMs present two challenges, however; the non-uniqueness of flux decomposition in terms of EFMs, and scaling of EFM analysis in large networks. Here, I address both computational problems and show how EFM analysis can be used to characterize the flow of metabolites and their atomic constituents in large-scale metabolic networks. In the first part of this thesis, I develop a biophysically-motivated method enforcing a Markovian constraint to uniquely decompose fluxes onto EFMs in strictly-unimolecular reaction networks. I refer to this method as the cycle-history Markov chain (CHMC) and prove its correctness with both discrete- and continuous-time Markov chains. Using the CHMC, I address biophysical questions regarding the distribution of pathway fluxes in a unimolecular, sphingolipid kinetic model of healthy and Alzheimer's disease patients. My statistical analyses of EFM weights support a dominant pathway flux hypothesis, whereby the majority of network fluxes are explained by a small subset of highly active EFMs. In my final aim, I generalize my Markov chain method to any type of metabolic network, including those with multispecies reactions. I do this through my proposal of atomic elementary flux modes (AEFMs) which explain the minimal, steady state flow of indivisible atoms in a metabolic network. By constraining atomic movements with atom mapping predictions, I show that AEFMs, unlike EFMs, can be enumerated in five large-scale metabolic networks by means of an atomic cycle-history Markov chain (ACHMC). In a subsequent analysis of a single set of inferred fluxes in a human liver cancer cell line (HepG2), I further show that glutamine-derived carbon AEFMs exhibit pathway flux dominance, with the most active AEFMs corresponding to well-known metabolic subsystems and the recently discovered non-canonical tricarboxylic acid (TCA) cycle. Altogether, my (atomic) cycle-history Markov chain ((A)CHMC) methods address fundamental challenges in EFM analysis and showcase the potential of both EFMs and AEFMs to further our understanding of cellular metabolism.

Résumé

Le métabolisme cellulaire est un process dynamique régulant la production et la consommation de métabolites. Ces capacités de réactions et/ou les flux métaboliques sont régulés à différemment niveau cellulaire par les gènes, les transcrits, les protéines et les métabolites. Les avancées technologiques et les nouvelles méthodes computationnelles permettent d'estimer de mieux en mieux les flux homéostatiques. Dans cette thèse, je me suis concentré sur l'avancement de certaines méthodes pour analyser les données de flux homéostatiques utilisant les voies signalétiques minimales connues comme mode de flux élémentaire (MFE). Puisqu'un ensemble arbitraire de flux homéostatique peut être reconstruit par une combinaison linéaire positive des MFE, ces voies peuvent être considérées comme des unités fonctionnelles de flux à l'intérieur et à travers un réseau de flux métaboliques. Travailler avec des MFE présente toutefois deux difficultés : le caractère non unique de la décomposition des flux en termes de MFE et la mise à l'échelle des flux des grands réseaux. Ici, j'aborde les deux problèmes de calcul et montre comment cette analyse peut être utilisée pour caractériser les flux de métabolites et leurs constituants atomiques dans les systèmes métaboliques à grande échelle. Dans la première partie de cette thèse, je développe une méthode empruntée de la biophysique qui applique une contrainte markovienne pour décomposer de manière unique les flux sur les MFE dans les réseaux de réactions strictement unimoléculaires. Cette méthode est appelée CHMC et ce travail prouve sa justesse avec des chaînes de temps discret et continu de Markov. En utilisant la CHMC, j'aborde des questions biophysiques concernant la distribution des flux des voies dans le modèle cinétique unimoléculaire de sphingolipides de patients sains et de patients atteints de la maladie d'Alzheimer. Mes analyses statistiques des poids MFE soutiennent l'hypothèse d'un flux de voie dominant, par lequel la majorité des flux du réseau sont expliqués par un petit sous-ensemble d'MFE très actifs. Dans mon objectif final, je généralise ma méthode de chaîne de Markov à tout type de réseau métabolique, y compris ceux avec des réactions multi-espèces. Je le fais en proposant des mode de flux élémentaire atomiques (MFEA) qui expliquent le flux minimal et constant d'atomes indivisibles dans un réseau métabolique. En contraignant les mouvements atomiques avec des prédictions de cartographie atomique, je montre que les MFEAs, contrairement aux MFEs, peuvent être dénombrés dans cinq réseaux métaboliques à grande échelle au moyen d'une cycle-history Markov chain atomique ACHMC. Dans une analyse secondaire d'un seul ensemble de flux déduits dans une HepG2, je montre en outre que les MFEA de carbone dérivé de la glutamine présentent une dominance du flux de la voie avec les MFEA les plus actifs correspondant à des sous-systèmes métaboliques bien connus et au cycle TCA non canonique récemment découvert. Dans l'ensemble, mes méthodes (A)CHMC abordent des défis fondamentaux de l'analyse MFE et démontrent le potentiel des MFE et des MFEA pour approfondir notre compréhension du métabolisme cellulaire.

Acknowledgements

My PhD journey was made possible by a number of individuals who supported me during every stage of my degree. First, I would like to thank my supervisor Dr. Theodore (Ted) J. Perkins for his mentorship and guidance throughout the years. Since my Bachelors, he has taught me nearly everything I know on how to be an outstanding scientist. It is no surprise that many students who work with Ted end up staying for one or more graduate degrees. During the first year of my PhD, I recall him telling me to seek out truly unknown and cutting edge problems to work on. He then gave me the freedom and responsibility to develop my very own PhD project and I am truly grateful for this experience. I am also especially grateful to him for funding me on a number of conferences during my PhD which have contributed immensely to my professional development across the computational and biological fields that my work intersects. Something he may not know (until now) is that he is also, in large part, responsible for the rather excellent quality of my sleep. When Ted's intuition says something is possible, I can rest easy knowing that the challenges faced today will be overcome tomorrow.

I would also like to thank my mother for her immense support during my PhD, and a number of friends and colleagues within the University of Ottawa and the Ottawa Hospital Research Institute. There are too many individuals to name; however, I would like to start with Christopher J. Porter, Gareth Palidwor, and Asma Bankapur, who form the bioinformatics core facility, and are always around to support students in Ted's lab. Another thank you goes out to lab alumni including Soroush Fard for his sharp wit, Aseel Awdeh for our hikes in Gatineau park, and Renad Al-Ghazhawi for joining me at nearly every graduate event and two conferences. Many thanks goes to my brilliant colleagues around the department who helped me transition to working on site after years of isolation working from home due to the COVID19 pandemic. You shared your wisdom with me, adopted me into your lab circles, and made me laugh every day. Thank you to the proteomics core facility (Andrew Macklin); StemCore (Damien Carragher, Shahriar Sheikholeslami, and Fernando Ortiz), *all* members of the Rudnicki lab/Satellos (notably Bahareh Hekmatnejad, Corentin Guilhot, Shanti Rayagiri) and Stanford lab (notably Mzwanele Ngubo, Fereshteh Moradi, Alberto Camacho-Magallanes, Julien Yockell-Lievre), Harper lab (Luke Kennedy, Michel N. Kanaan, Claire Fong-McMaster, Dhanuddara Mohottalage), Lee lab (Danny Farhat), Couture lab (Hossein Davarinejad), Lavallée-Adam lab (Dallas J. Nygard, Rachel Nadeau, Aarthie Senathirajah, Kyle Tomaro), Stintzi lab (Peter Dobranowski), Sun lab (Katrina Madden), Côté lab (Redaet Daniel), Graber lab (Sean Stephenson), Power lab (Dawson B. H. Livingstone), Djaoud lab (Yeganeh Almasi), and Chan lab (Mahanish J. Thapa). I acknowledge the support of many these professors as well for their mentorship during my times of need. A special thank you also goes out to my thesis advisory committee members Drs. Yan Burelle,

Mireille Khacho for their enthusiasm and bravery advising me, their first computational student, and Dr. Mads Kærn for his encouragement on communicating flux analysis to biologists.

It would be remiss to forget the group of international colleagues I have met over the scientific conference circuits and the perennial friendships formed. I cannot name everyone but would like to acknowledge Salvador E. Caoili, Carl Munoz, Igor Martayan, Ali S. Imami, Kosuke Shiraishi, Eric Cassens, Jona Borges, Song Cao, and Rita R. Manuel. You made each conference I attended a special one.

Lastly, I acknowledge those who shall not be named at the University of Ottawa who unwittingly altered the direction of my research career. This thesis serves as a testament to those who have or will encounter challenges during their PhD. Persevere for you are not alone.

This work was supported by several institutes and agencies which gave me the financial security to pursue this project. These included travel funding from the Ottawa Hospital Research Institute and University of Ottawa Faculty of Medicine; and computing resources from Compute Canada (rest in peace) and the Digital Research Alliance of Canada. I also acknowledge the support of the Canadian federal and provincial government through the Natural Sciences and Research Council of Canada Collaborative Research and Training Experience-Metabolomics Advanced Training and International Exchange (NSERC-CREATE MATRIX) program, Alexander Graham Bell Canada Graduate Scholarship-Doctoral program (NSERC CGS-D), and Ontario Graduate Scholarship (OGS).

Contents

List of Figures

List of Tables

List of Abbreviations

β-GC	beta-glucosylceramidase
(A)CHMC	(atomic) cycle-history Markov chain
(A)EFM	(atomic) elementary flux mode
SLC25A11	solute carrier family 25 member 11
SLC25A12	solute carrier family 25 member 12
SLC25A13	solute carrier family 25 member 13
SLC25A1	solute carrier family 25 member 1
2THD	hexadecanal
6PG	6-phosphogluconate
ACDase	acid ceramidase
ACHMC	atomic cycle-history Markov chain
ADP	adenosine diphosphate
AEFM	atomic elementary flux mode
akg	alpha-ketoglutarate
ALE	adaptive laboratory evolution
AlkCDase2-3	alkaline ceramidase 2-3
AlkCDase3	alkaline ceramidase 3
aSMase	acid sphingomyelinase
asp	aspartate
ATP	adenosine triphosphate
BFGS	Broyden–Fletcher–Goldfarb–Shanno
BiGG	Biochemical Genetic and Genomic
BWT	Burrows-Wheeler Transform
C	cytosol
C1	one carbon
C1P	ceramide-1-phosphate
C2C12	immortalized mouse myoblast
CDF	cumulative distribution function
CER	ceramide
CERK	ceramide kinase

CerS	ceramide synthase
CERT	ceramide transport protein
CFP	carbon flux path
ChIP-seq	chromatin immunoprecipitation followed by sequencing
CHMC	cycle-history Markov chain
cit	citrate
CO_2	carbon dioxide
CoA	coenzyme A
CTMC	continuous-time Markov chain
DHAP	dihydroxyacetone phosphate
DMEM	Dulbecco's Modified Eagle Medium
DNA	deoxyribonucleic acid
DNF	did not finish
DTMC	discrete-time Markov chain
E. coli core	*E. coli*
ECaM	elementary carbon mode
ECAR	extracellular acidification rate
ECM	elementary conversion mode
EFM	elementary flux mode
EMP	Embden-Meyerhof-Parnas
EP	extreme pathway
ER	endoplasmic reticulum
F6P	fructose 6-phosphate
FAPP2	four-phosphate adaptor protein 2
FBA	flux balance analysis
FDP	fructose 1,6-bisphosphate
fum	fumarate
FVA	flux variability analysis
G3P	glyceraldehyde 3-phosphate
G6P	glucose 6-phosphate
GA	golgi apparatus
GACF	golgo apparatus (cytoplasmic face)
GCS	glucosylceramide synthase
GED	Gnd-Entner-Doudoroff
GEM	genome-scale metabolic model
Glc	glucose
gln	glutamine
glu	glutamate
GluCER	glucosylceramide
GPR	gene-protein-reaction

GSL	glycosphingolipid
H_2O	water
HepG2	human liver cancer cell line
HPLC	high-performance liquid chromatography
iAB RBC 283	red blood cell
icit	isocitrate
iIT341	*H. pylori*
IM	inner membrane
iSB619	*S. aureus*
KEGG	Kyoto encyclopedia of genes and genomes
KS	Kolmogorov-Smirnov
L	lysosome
L2 norm	Euclidean norm
LacCer	lactosylceramide
LacCerS	lactosylceramide synthase
LP	linear programming
lSPA	shortest pathway activity by linear programming
M	mitochondria
m/z	mass over charge ratio
mal	malate
max	maximum
MC	Markov chain
MCA	metabolic control analysis
MCS	minimal cut set
MFE	mode de flux élémentaire
MFEA	mode de flux élémentaire atomiques
MIDAS	mass spectrometry integrated with equilibrium dialysis for the discovery of allostery systematically
milAP	active pathways by mixed integer linear programming
MILP	mixed-integer linear programming
min	minimum
MLE	maximum likelihood estimation
MRM	multiple reaction monitoring
MS	mass spectrometer
N	nucleus
NAD^+	nicotinamide adenine dinucleotide (oxidized form)
NADH	nicotinamide adenine dinucleotide (reduced form)
NCDase	neutral ceramidase

NSCLC	non-small cell lung cancer
nSMase2	neutral sphingomyelinase 2
oaa	oxaloacetate
OCR	oxygen consumption rate
ODE	ordinary differential equation
OM	outer membrane
P-P	probability-probability
P_i	inorganic phosphate
PAP2a-b	phosphatidate phosphohydrolase 2a-b
PAP2a-b-c	phosphatidate phosphohydrolase 2a-b-c
PDF	probability density function
PhET	phosphoethanolamine
PPP	pentose phosphate pathway
Q-Q	quantile-quantile
QP	quadratic programming
qSPA	shortest pathway activity by quadratic programming
RNA	ribonucleic acid
S1P	sphingosine-1-phosphate
SBML	Systems Biology Markup Language
SITA	stable isotope tracing analysis
SK 1-2	sphingosine kinase 1-2
SK1	sphingosine kinase 1
SK2	sphingosine kinase 2
SM	sphingomyelin
SMILES	simplified molecular input entry system
SMS1	sphingomyelin synthase 1
SPH	sphingosine
SPL1	sphingosine-1-phosphate lyase
SPP1	sphingosine-1-phosphate phosphatase 1
SPP1-2	sphingosine-1-phosphate phosphatase 1-2
SRM	selected reaction monitoring
succ	succinate
succoa	succinyl-CoA
TCA	tricarboxylic acid
THF	tetrahydrofolate
TPM	transcripts per million
YV	yield vector

A process cannot be understood by stopping it. Understanding must move with the flow of the process, must join it and flow with it.

—Frank Herbert, *Dune*

1

Introduction

1.1 Overview

Metabolism is defined as the collection of biochemical reactions within an organism to sustain life. These catabolic and anabolic reactions are constantly remodelling nutrient sources to produce endogenous metabolites necessary for cellular survival and proliferation. Thus far, over 100 metabolic models of bacterial, yeast, and mammalian cells have been published in the Biochemical Genetic and Genomic (BiGG) database alone,[1,2] consisting of thousands of known metabolites, reactions, and genes identified from over a century of biochemical studies.

It cannot be stressed enough how remarkable the field of metabolism has advanced since the discovery of the enzyme at the turn of the 20th century.[3] This period marked the birth of modern biochemistry and the burgeoning field of metabolism. It was during this time that the metabolic pathways of central carbon metabolism as we know today

were elucidated. As chemists moved into biochemistry, new enzymes and metabolic intermediates were discovered, leading to a comprehensive understanding of carbohydrate metabolism, photosynthesis, and oxidative phosphorylation among other pathways. It also cannot be understated how these metabolic pathways were discovered using analytical chemistry techniques considered crude by modern standards. Using colorimetry and paper chromatography, for example, HA Krebs and WA Johnson discovered the tricarboxylic acid (TCA) cycle in 1937.[4–6] Shortly after in the 1940s the Embden-Meyerhof-Parnas (EMP) (glycolytic) pathway was full characterized[7], followed by the pentose phosphate pathway (PPP) by Horecker et al. in 1951.[8,9] With the discovery of ^{14}C by Ruben and Kamen, radiolabelling techniques were instrumental in identifying the Calvin-Benson-Bassham cycle in 1954.[11,12] Subsequent manometric experiments were crucial in characterizing aspects of mitochondria metabolism, such as the discovery that nicotinamide adenine dinucleotide (reduced form) (NADH) was compartmentalized between the cytosol and mitochondria[13] and the malate-aspartate cycle[14] among other NADH shuttle systems.[15] These pathways are now assembled into metabolic maps[16] and genome-scale metabolic models (GEMs) curated for individual organisms and cells, reflecting over a century of biochemical discoveries.

Since these discoveries, many questions in metabolism have shifted from the identification of individual pathways to the interactions between pathways to obtain a deeper understanding of cellular metabolism. This systems biology approach has been driven, in large part, by advancements in high-throughput sequencing and other omics-based approaches to identify and quantify the genes, transcripts, proteins, and metabolites within a single cell.[17,18] Increasingly sophisticated experimental techniques have revealed even more complex interactions within and across data types. Examples include ChIP-seq to identify DNA-protein interactions,[19] Hi-C to study the 3D conformation of DNA,[20] BioID to study protein-protein interactions,[21] and more recently MIDAS to study protein-metabolite interactions.[22]

In parallel to these omics advancements, systems biology has also been driven by mathematical theories and mechanistic frameworks that have led to biological discoveries and reshaped the field of metabolism. These methods formalize our existing biochemical knowledge and their complex interactions to understand how they give rise to global cellular properties.[23–26] One systemic property of cellular metabolism is metabolic flux which is defined as the rate of biochemical reactions. Perhaps the most famous framework to study flux is metabolic control analysis (MCA), which emerged independently in the 1970s by Kacser and Burns,[27,28] and Heinrich and Rapoport.[29,30] MCA was developed as a means to quantify the steady state interactions between fluxes, enzymes, metabolites in response to various perturbations. A major discovery of MCA was that all enzymes exert some level of control on flux through a metabolic pathway, disproving the prevailing notion that pathway fluxes were controlled at rate-limiting steps by "pacemaker enzymes."*[32,33] Variations of this concept remain today with emerging ideas on metabolic reprogramming and adaptability to explain changes in metabolic (dys)regulation.[34] Another important mathematical method is flux balance analysis (FBA) which is a framework to understand the distribution of steady state fluxes under simulated or experimentally-derived biological constraints.[35,36] Altogether, these flux analysis methods are becoming increasingly relevant in today's data-rich era of biology to understand how metabolism changes as a function of the genome, transcriptome, proteome, and metabolome.

One specific type of flux analysis, and focus of this thesis, is elementary flux mode (EFM) analysis. The concept of EFMs was first proposed in 1994 by S Schuster and Hilgetag[37] as a method to study pathway fluxes in stoichiometric models of metabolism.

*It has taken many years to dispel the myth of the rate-limiting step by convincing biochemists that mathematical models can drive biological discoveries. I sympathize with Fell and Sauro who voiced these frustrations nearly 40 years ago regarding the power of MCA to understand cellular regulation who said, "Whilst this seems a legitimate and fruitful use of the detailed knowledge of biochemical mechanisms gained empirically over the years, it is regarded with deep suspicion by many biochemists, who seem to feel a misplaced confidence in their ability to achieve an intuitive and qualitative understanding of these non-linear multivariable problems without putting it to the rigorous and quantitative test implied by simulation."[31(p556)]

An EFM can be defined as a minimal set of reactions that can sustain steady state flux (see Chapter 2 for a more mathematical description). This implies that the sequence of reactions balances all internal metabolite stoichiometries and that the capacity to maintain steady state flow is lost if any reaction is deleted. Put another way, an EFM is a steady state pathway that does not contain another steady state pathway.

EFMs have several useful properties that make them ideal for studying the structure of metabolic networks and the distribution of pathway fluxes. For one, EFMs are highly intuitive pathways to understand and visualize. In a metabolic network consisting strictly of unimolecular reactions, EFMs correspond to either simple paths explaining the movement of a metabolite from source to sink metabolites (Figure 1.1a), or simple cycles which are biologically analogous to substrate cycles (Figure 1.1b). When multispecies reactions are present, this notion of simple path and cycle extends to include branching and converging pathways carrying steady state flux (Figure 1.1c). Since the removal of any reaction in an EFM abolishes steady state flux, EFM can therefore be viewed as the possible mechanisms for carrying flux within a network. EFMs have been analysed to characterize the robustness of metabolic networks,[25,38–40] and to guide the rational design of cells optimized for growth or the production of desirable metabolic compounds.[41–43]

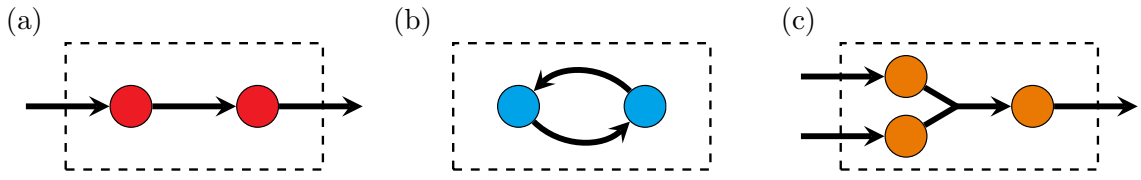


Figure 1.1: EFMs corresponding to an (a) simple path. (b) simple cycle. (c) Path involving a multispecies reaction.

Another important property of EFMs is that there exists a finite number of these pathways characterizing a given network which are unique to a scalar multiple. This property may sound trivial but consider how there are infinite ways for a metabolite to move through most metabolic networks. A metabolite could cycle indefinitely between

internal metabolites in a closed-loop network or re-enter the network repeatedly in an open-loop network. However, a metabolite moving within the same substrate cycle once, twice, or a million times would still constitute a flow through that corresponding EFM with a weight of one, two, or a million. From this realization, one can observe that the steady state fluxes in any network, no matter how complex, can be written as a positive, linear combination of their EFMs. This final property enables one to represent the complex distribution of steady state fluxes in a network in terms of functional units of pathway flux. One could envision applying EFMs to address questions regarding the structures of large-scale metabolic networks. With the rise of methods to estimate and infer metabolic fluxes, one could also imagine applying EFM analyses to quantify the distribution of pathway fluxes within and across biological conditions.

1.2 Problem statement

These properties of EFMs make them suitable for addressing a fundamental biological problem regarding the nature of pathway fluxes in a metabolic flux network. However, the practicality of conducting these EFM analyses leads to two subsequent computational problems that remain unsolved in EFM literature to this day.

Understanding the flow of metabolites from steady state flux data

A fundamental question of flux analysis is inferring the movement of metabolites in a network. This question is interesting from a biophysical standpoint to understand the dynamics of a metabolite remodelling in a complex network of reactions, but also important for understanding how flux is distributed across metabolic pathways. In recent years, this question has been approached statistically using pathway analysis methods leveraging widely-available transcriptomics and metabolomics datasets. Methods

such as over-representation analysis (ORA) employ a hypergeometric test to identify metabolic pathways with a greater-than-expected number of differentially expressed genes or differentially abundant metabolites.[44,45]. However these methods are highly sensitive to metabolic pathway definitions, P-value thresholds, and other statistical significance scores.[46,47] More pressingly, these methods cannot provide a mechanistic explanation for why a given metabolic pathway is active. They cannot distinguish between differentially active or inactive pathways from abundance data (since increased substrate levels may be counteracted by insufficient enzyme levels), identify which reactions within that pathway are differentially active, or explain the interactions between differentially active pathways. These problems persist among more sophisticated pathway analysis methods such as functional class scoring[48,49] or topology-based methods.[50,51] Hence, new flux analysis methods are required to understand how metabolites transit in a network both within and across metabolic pathways.

1.2.1 Decomposing fluxes onto EFMs

The biological problem posed in the previous section can be transformed into a computational one regarding the analysis of EFMs. I refer to this as the EFM flux decomposition problem which is to determine how much flux is travelling along each EFM given a metabolic flux network under steady state. The challenge, however, is that decomposing steady state fluxes onto EFMs is generally not unique. There are typically more EFMs than linearly independent fluxes in a network, leading to an infinite set of EFM weights explaining the observed network fluxes.

This problem of uniquely decomposing fluxes onto EFMs has limited its applications for studying the structure of metabolic networks and understanding the distribution of pathway fluxes. Numerous mathematical optimization strategies using linear programming (LP), quadratic programming (QP), and mixed-integer linear programming

(MILP) have been proposed to explain a metabolic flux network in terms of a single set of EFM weights. However, these mathematical optimization methods are poorly justified from a biophysical standpoint. First, a metabolite has no memory of the reactions leading to its current state, nor knowledge of an EFM and the sequence of reactions it must undergo to complete that pathway. Second, objective functions proposed in literature are based on notions of parsimony that are generally biophysically implausible. For example, some groups have proposed pathway fluxes are explained by the fewest number of EFMs or EFMs with the fewest number of reactions,[52–54] while others have proposed objective functions that do the exact opposite.[55,56] Finally, even these objective functions are not guaranteed to constrain the solution space to a single set of EFM weights. Without a unique flux decomposition, one cannot disentangle whether changes in EFM weights are an artifact of the chosen objective function and solver, or reflect genuine metabolic pathway activity.

1.2.2 Enumerating EFMs in large-scale metabolic networks

Another major challenge in EFM literature is the complete enumeration of all steady state pathways in a metabolic network, and has been the historical focus of EFM research. This remains an ongoing problem in literature since the number of EFMs can be unfathomably large. To understand why there are so many EFMs, consider how metabolic networks exhibit a high degree of connectivity between metabolites and often through multispecies reactions. This leads to a combinatorial explosion of EFMs balancing different sets of internal metabolite stoichiometries. The result is that EFM enumeration is computationally infeasible for large-scale metabolic networks after decades of algorithmic advancements.[57–62] Empirical studies using these tools have demonstrated over a million EFMs in small-scale metabolic models with fewer than 100 metabolites and reactions.[60] The synthetic, minial cell *JCVI-syn3A*[63] consists of

304 metabolites and 338 reactions, and is predicted to contain over a trillion EFMs that would take years to enumerate with nearly one thousand threads running in parallel.[61] EFM enumeration is therefore computationally infeasible for most GEMs with even the question of determining the complexity of enumerating all EFMs in a network remaining unclear.[64,65]

This computational infeasibility of EFM enumeration is a major obstacle to EFM-based flux analysis. Without complete knowledge of all network EFMs, both structural and flux analyses of these pathways are meaningless. EFMs are a structural property of the network and incomplete knowledge of all pathways can bias structural analyses involving EFMs. Similarly, analysis of EFM weights may be prone to bias if biologically relevant EFMs are missing. While many groups have proposed methods to enumerate subsets of EFMs satisfying specific biological properties, contextualizing the significance of these pathways at a structural or pathway flux level is only possible when all EFMs are known.

1.3 Thesis statement

Metabolic fluxes are a rich source of information describing metabolic dynamics and are becoming increasingly relevant as methods improve to estimate them advance. However, few methods exist today to analyse flux data. EFMs are an elegant mathematical framework to study metabolic flux networks, yet their potential for analysing large-scale networks has been hindered by the longstanding EFM flux decomposition problem. For decades, decomposing fluxes onto EFMs has been approached by mathematical optimization methods which assume individual metabolites are predestined to travel along a given EFM based on biophysically implausible notions of parsimony. I argue that a new perspective to modelling pathway fluxes is required to advance systems biology and unravel the complexity of cellular metabolism at the genome scale.

1.4 Contributions

This work aims to solve longstanding problems in EFM literature to bring forth a new generation of methods and perspectives to analyse steady state metabolic flux data. My main contributions are as follows:

1. Developing a probabilistic, Markov chain model to solve the EFM flux decomposition problem in unimolecular metabolic flux networks.

2. Framing this probabilistic model in terms of existing stochastic chemical kinetic literature.

3. Addressing the nature of pathway flux distributions within and across biological conditions.

4. Proposing a new type of steady state pathway to characterize the flow of atomic constituents in any type of metabolic flux network.

5. Developing a computational pipeline to both enumerate these pathways and uniquely decompose steady state fluxes onto them under a Markovian constraint.

6. Demonstrating that these computations are feasible on large-scale metabolic networks with existing consumer-grade computing resources.

7. Predicting the flow of source metabolite carbons in a human liver cancer cell line (HepG2) and comprehensively analysing the dominant pathways explaining glutamine carbon remodelling.

8. Developing and documenting an open-source software package that implements these novel algorithms for EFM analysis.

1.5 Thesis organization

In **Chapter 2**, I review methods in literature to generate and analyse metabolic flux data. The following **Chapters 3–7** detail my main contributions to advance flux analysis through the concept of EFMs. In **Chapter 3**, I revisit the EFM flux decomposition problem and show how existing optimization-based methods proposed in literature are biophysically implausible and not guaranteed to uniquely identify EFM weights. These observations lead to my proposal of a certain discrete-time Markov chain model, known as the cycle-history Markov chain (CHMC), to uniquely decompose steady state fluxes onto EFMs for strictly unimolecular flux networks. In **Chapter 4**, I show how to derive my Markov chain method in continuous-time, starting from a system of mass-action differential equations, and prove that both approaches lead to identical EFM probabilities and weights. Using my EFM weight assignment method, I characterize the highly uneven distribution of EFM weights within and across biological conditions in **Chapter 5**, supporting a *dominant pathway flux hypothesis*. In **Chapters 6 and 7**, I address both the generalized EFM flux decomposition problem and EFM enumeration problem through my introduction of atomic elementary flux modes (AEFMs). I then describe a computational pipeline for enumerating and uniquely identifying AEFM weights in large-scale, multispecies reaction networks. I show that these pathways, unlike standard EFMs, can be computed in GEMs, and are arguably more biologically meaningful to interpret through a structural analysis of five GEMs and AEFM weight analysis of a HepG2 cell line. A summary of my work, limitations, and future directions are finally presented in **Chapter 8**.

2

Literature review

In this chapter, I review the literature on methods to analyse metabolic flux networks. I first begin by explaining how metabolic networks are constructed at the genome-level before discussing methods to estimate steady state fluxes in these networks. I then describe computational methods to analyse these networks with and without knowledge of their corresponding steady state fluxes. I then provide some background on simplified molecular input entry system (SMILES) strings before reviewing some basic Markov chain theory required to understand my contributions in Chapters 3–7.

2.1 Genome-scale metabolic models (GEMs)

Genome-scale metabolic models (GEMs) refer to the reconstruction of an organism's metabolism defined by a system of metabolites and corresponding reactions. The size and scope of these models vary and are associated with both large-scale metabolic net-

works (e.g. SA Becker and Palsson[66]) and smaller scale ones defined by a few metabolic subsystems (e.g. Orth et al.[36]). Historically, these models were manually constructed by curating known metabolites, reactions, and genes characterized for a given organism.[67] These information were systematically organized to link different biochemical data in addition to known biochemical constraints. Examples of the former are gene-protein-reaction (GPR) rules[68] which encode the relationship between a gene, its gene product, and how that gene product drives flux through a reaction by itself as an enzyme/enzyme isoform, or with the cooperation of other proteins as an enzymatic subunit. Examples of biochemical constraints include the stoichiometric relations between participating metabolites that lead to atomic mass conservation within reactions.

The advent of high-throughput sequencing and other omic technologies has led to an explosion of data to reconstruct GEMs. Today, there exist many algorithms and software suites to automate the GEM construction process. Notable programs include RAVEN[69], COBRA[70], and KBase[71] which has been used to draft over 80,000 GEMs; however, many more programs have been developed since then.[72] The modern approach to constructing a GEM generally involves acquiring large-scale sequencing data specific to an organism of interest. These reads are mapped to a reference genome, or a homologous species if one is unavailable, to identify the metabolic enzymes from genome annotation databases.[73,74] These enzymes are then mapped to reaction-specific databases to identify the corresponding substrates and products (e.g. KEGG[75], ModelSeed[76], TCDB[77]). The result is a draft GEM that may require multiple pre-processing steps, also incorporated in many GEM reconstruction software programs, to finalize the metabolic model.

A draft GEM may require one of several pre-processing steps after initial assembly from a reference genome. For one, the resulting list of metabolites and reactions identified from sequencing data may not form a singular metabolic network. Not all enzymes may be identified or protein transporters that move substrates across subcellular com-

partments. The reference genome may also be incomplete with uncharacterized genes with no known biological function. These problems can lead to unconnected metabolic subnetworks that are orphaned from one another due to missing reactions and dead-end metabolites that cannot maintain steady state flux. These problems reflect incomplete metabolic knowledge of the organism of interest and necessitate gap-filling algorithms[78] and expert curation to ensure a strongly-connected network where steady state flux can be routed through all participating metabolites.

2.2 Generating metabolic flux data

Once the network structure is defined, the metabolic flux along each reaction may be measured or estimated from a host of experimental and bioinformatic techniques. These fluxes are often estimated under metabolic steady state, where one assumes zero net change in metabolite concentrations over time, so the total metabolic fluxes producing and consuming each metabolite are balanced. While experimental methods are generally considered more accurate than computational ones, both methods are used in complementary ways. In the following sections, I describe common experimental and computational methods to estimate and infer steady state fluxes. I also show how experimentally deduced fluxes can be used as a ground truth to validate computational models of flux or serve as model inputs to refine flux inferences.

2.2.1 Experimentally deducing metabolic fluxes

In this section, I cover the following experimental techniques to estimate steady state fluxes and provide several examples of their use case in literature. These strategies include (i) time-resolved exometabolomics, (ii) radiolabelling, and (iii) stable-isotope tracing analysis.

Time-resolved exometabolomics

Exometabolomics refers to the identification and measurement of exogenous metabolites present in cell culture media. This technique is therefore useful for identifying source and sink metabolites that are exchanged between cells and the media. These source and sink fluxes are estimated by quantifying the abundance of metabolites in the media over time, accounting for the number of cells growing in culture. The external fluxes are subsequently computed by fitting the time-dependent abundances to differential equations describing the extracellular metabolite fluxes released into the media or taken into the cells. Nilsson et al., for example, used this approach to measure 23 extracellular fluxes in human liver cancer cell line (HepG2) cells corresponding to amino acids, glucose, lactate, and pyruvate.

Several techniques are commonly used to estimate the metabolite abundances in exometabolomic studies. Common protein targets can be quantified by colorimetric assays (e.g. *Pierce BCA protein assay kit* from ThermoScientific). Other types of metabolites can be identified by high-performance liquid chromatography (HPLC) and quantified by their refractive index or ultraviolet light absorbance with the aid of calibration curves. More complex methods include targeted mass spectrometry data acquisition methods such as selected reaction monitoring (SRM) or multiple reaction monitoring (MRM) to identify compounds based on their mass over charge ratio (m/z) and quantify their abundances using internal standards.[80] A commercial platform to measure specific extracellular fluxes is the Agilent Seahorse analyser. This platform specifically assesses mitochondrial respiration and glycolysis through oxygen consumption rate (OCR) and extracellular acidification rate (ECAR), respectively, by measuring the dissolved oxygen and free protons in the cell culture media.

Radiolabelling

In contrast with exometabolomics, more specialized tracer-based techniques are required to estimate intracellular fluxes. These tracers are small molecules that are identifiable by their radioactive signature and can hence be traced through a series of reactions. These molecules contain unstable combinations of protons and neutrons that cause the molecule to decay and release radiation. In the 20th century, ^{14}C tracers were extensively used in radiolabelling studies to both identify and quantify flux through metabolic pathways. ^{14}C decays follows the following nuclear equation ${}^{14}_{6}C \rightarrow {}^{14}_{7}N + \beta^-$, where β^- is a high-energy, high-speed electron.[81] The carbon isotope thus decays into a stable nitrogen isotope releasing radioactive energy in the process. A benefit of ^{14}C over shorter-lived radioisotopes is its very long half-life of 5730 $\pm$ 40 years,[82] allowing it to be traced over the course of hours or even days.

In the mid 20th century, radiolabelling studies were primarily used to identify metabolic pathways.[11,12] These were performed by culturing cells in radiolabelled compounds and observing the radioactive metabolites in downstream pathways. For example, an early 1960s paper by Ap Rees and Beevers used labelled pyruvate and glucose to quantify their breakdown in carrot tissue through the pentose phosphate pathway (PPP) and Embden-Meyerhof-Parnas (EMP) pathway.[83] In this experiment and many others, paper chromatography was used to separate complex metabolite mixtures. Autoradiographs were then used to detect and quantify radioactive metabolites from these paper chromatographs. This technique involves exposing the radioactive compounds to a thin layer of silver halide which ionizes in close proximity to the energy released by radioisotopes. The result is a silver halide precipitate that appears as a black spot visible to the naked eye.[84] Exposure could take weeks to months due to the long half-life of ^{14}C.[84] These spots were then excised and radioactivity quantified by oxidizing the compounds and measuring the quantity of precipitates or emitted photons from the β^- particles using a liquid scintillation counter.[12,83] Today, radiolabelled tracers are

still quantified by liquid scintillating counters[85] and by more advanced accelerator mass spectrometers[86] which directly count the number of radioactive atoms rather than the scintillations emitted from decay events.

Stable isotope tracing

In recent years, stable isotope tracing has become the standard experimental technique for experimentally deducing metabolic fluxes. As the name suggests, stable isotope tracers refer to chemical compounds with a stable configuration of neutrons. ^{13}C, for example, is a stable isotope of carbon 12 which contains 6 protons and neutrons. In contrast with radioisotopes, stable isotopes do not decay and therefore do not release radiation. This technique offers several advantages over radiolabelled tracers including higher throughput flux estimates, rapid sample preparation, and possibility to analyse multiple and/or distinct metabolic pathways using a combination of stable isotope tracers.[72]

Mass spectrometer (MS)-based technologies are the primary means to trace the incorporation of stable isotopically-labelled substrates into endogenous metabolites. At a high level, MS is a platform with the capability to (i) record the m/z ratio of charged molecules, (ii) select for molecules with given a m/z, and (iii) fragment molecules into smaller constituents. As a simple example, one could imagine identifying an endogenous metabolite based on its characteristic m/z (M) and quantifying that metabolite based on the number of detected ions with that m/z. One could then measure whether that endogenous metabolite contained additional neutrons (e.g. $M+1$, $M+2$) donated by the isotopic tracer moving through the metabolic network. The combination of these three capabilities leads to distinct operational protocols referred to as MS data acquisition modes. Each mode has its advantage and disadvantage, but modes are typically classified into "targeted" and "untargeted" approaches. Broadly speaking,

targeted mass spectrometry excels at quantifying metabolite levels but require *a priori* structural knowledge of the metabolites of interest. This is the main approach used in stable isotope tracing studies to investigate metabolic flux in well-studied pathways such as central carbon metabolism[87] and within subcellular compartments.[88] Untargeted MS approaches are much less common since they trade quantification accuracy for the detection of all molecules in the biological system. Computing the ratio of all labelled-versus-unlabelled metabolites from untargeted MS is sometimes used to obtain a semi-quantitative understanding of metabolic flux.[89]

It should be highlighted that the versatility of targeted stable isotope tracing studies involves a great deal of computational complexity not found in untargeted or even radiolabelling strategies. In the simple example described above, I explained how the flux of a stable isotopically-labelled compound could hypothetically be traced by quantifying the number of downstream metabolites that incorporated those additional neutrons. In reality, however, identifying metabolic fluxes is much more complicated due to the highly-connected nature of metabolic networks and the numerous ways that isotopes can incorporate into metabolites. Understanding this complexity requires introducing two concepts that formalize how isotopes can incorporate into endogenous metabolites. The first term is *isotopomer* which refers to a given compound with the same number of additional neutrons, but localized to different atoms within the molecule.[87] Isotopomers often arise in cyclic pathways, such as the tricarboxylic acid (TCA) cycle, where molecules can both inherit and relinquish isotopes at different atomic positions. The identification of mass isotopomer distributions is essential for measuring metabolic flux since this position-specific information is highly informative of the tracer movement. However, isotopomers cannot be identified by MS alone since molecules with the same number of additional neutrons in different configurations share the same m/z. Their identification requires an intimate understanding of the metabolic network structure and atomic movements across series of reactions to predict their labelling distributions.

A related term, not to be confused with the isotopomer, is the *isotopologue* which refers to the same chemical species but with a different configuration of neutrons.[90] Isotopologues occur naturally as about 99% of all carbon is the ^{12}C isotope and the quantity of detected metabolites with a given m/z must be corrected to account for this isotopic variation in nature.[91] Isotopologues also arise in complex metabolic pathways such as the TCA cycle. For example, citrate inherit two isotopes from uniformly labelled ^{13}C pyruvate during the first turn of the cycle and a total of three isotopes during the second turn. Analysis of both isotopologue and isotopomeric information is required to deconvolute targeted MS data to estimate the steady state fluxes in a metabolic network.[92]

2.2.2 Computationally inferring metabolic fluxes

Since reactions can occur very quickly, it can be challenging to observe a radioisotope or stable isotope tracer moving through a metabolic network. This problem has motivated two families of computational methods to infer fluxes from various types of biological data.

Flux balance analysis (FBA)

Flux balance analysis (FBA) is a versatile method to estimate a single set of steady state fluxes in a metabolic network. It is considered a stoichiometric technique since it operates on the structure of the metabolic network encoded as a stoichiometry matrix. This matrix S has m rows corresponding to the number of distinct network metabolites and n columns corresponding to the network reactions. The coefficients S_{ij} of the stoichiometric matrix represent the quantity of each metabolite $i = 1, 2, \ldots, m$ that is produced (positive coefficient) or consumed (negative coefficient) in reaction $j = 1, 2, \ldots, n$.

Given only the stoichiometric quantities involved in each reaction, FBA formulates a system of linear equations describing the mass balances for each metabolite

$$\frac{dX}{dt} = S \cdot v, \tag{2.1}$$

where X is a vector metabolite concentrations of length m and v is the flux vector of length n. A core assumption of FBA is that metabolic reactions occur faster than cellular growth or environmental changes, leading to metabolic fluxes that are under quasi-steady state.[35,93] This implies that the change in metabolite levels is zero and

$$\frac{dX}{dt} = 0 = S \cdot v. \tag{2.2}$$

Solving Equation (2.2) is equivalent to finding a set of steady state fluxes with no net change across all internal metabolite stoichiometries. However, there are often more fluxes (n) than mass balance constraints (m), leading to an underdetermined system of equations and an infinite set of fluxes that can satisfy steady state. This finding agrees with our biological intuition since, without enzyme-kinetic knowledge organism, a cell has an infinite number of ways to distribute fluxes through different combinations of reactions. More biological information or metabolic assumptions are required to constrain the solution space to a single set of steady state fluxes.

FBA incorporates three additional sets of constraints to identify a set of steady state fluxes that could feasibly describe the metabolic system in question. The first constraint takes the form of lower and upper flux boundaries (l_j and u_j) for reaction j in the stoichiometry matrix.[36] The boundaries for exchange reactions may be determined from exometabolomic studies, while those for internal reactions may be deduced from isotope tracing studies. * Another important constraint is that reaction thermodynamics are obeyed whereby irreversible reactions are never assigned negative fluxes implying flow

*A third omics option is discussed at the end of this section.

from products back to substrates. Combined with Equation (2.2), these two inequalities constrain the solution space of fluxes to a convex polyhedral cone[94] otherwise referred to as a flux cone.[95] The tip of this flux cone emerges from the origin, since no flux through all reactions is a feasible solution to Equation (2.2). As a cone, however, there are still many feasible sets of fluxes within this space that satisfy metabolic steady state.

Narrowing the flux cone even further requires specific knowledge of the metabolic phenotype to identify a single set of steady state fluxes describing the cell in question. This metabolic phenotype is encoded into the third and final FBA constraint known as the objective function which corresponds to a weighted set of reactions whose combined score Z is either maximized or minimized. In FBA studies of prokaryotes and yeast, a metabolic phenotype optimized for biomass production is often used since this metabolic behaviour often occurs in nature due to the evolutionary advantage conferred by fast cellular growth[96]. Indeed, adaptive laboratory evolutions (ALEs) studies[97] have shown that serial batch culturing bacteria leads to evolved strains with improved growth phenotypes.[98,99] Encoding this metabolic phenotype into the objective function requires defining an artificial biomass reaction which consumes various stoichiometric units of substrates to produce some stoichiometric unit of biomass. These stoichiometries may be experimentally deduced by quantifying the macromoleculer fractions of carbohydrates, nucleic acids, and amino acids within a cell culture.[100] Maximizing this biomass reaction thus leads to a single set of steady state fluxes optimized for cellular growth. Altogether, Equation (2.2) and these three constraints define the following linear programming (LP) problem:

$$\begin{aligned} &\text{maximize} && Z \\ &\text{subject to} && S \cdot v = 0 \\ &&& v_r \geq 0, \qquad \text{for } r \in \text{irreversible reactions} \\ &&& l_j \leq v_j \leq u_j, \quad j = 1, \ldots, n \end{aligned} \tag{2.3}$$

where $Z = c^{\intercal}v$ and c is the vector of reaction weights contributing to objective function. For example, an objective maximizing biomass growth would set $Z = v_{biomass}$. More complex non-linear objective functions may be specified and solved through other mathematical optimization methods such as quadratic programming (QP) or mixed-integer linear programming (MILP).

The modular constraint-based system for defining FBA problems makes it a highly flexible framework for understanding the distribution of steady state fluxes. First, this system of linear equations, flux boundaries, and objective function can be solved by LP which is fast and scalable to GEMs.[36] Metabolic hypotheses can also be tested *in silico* by deleting reactions or forcing specific levels of flux through certain reactions to understand how these perturbations change other network fluxes. Exchange flux boundaries can also be modified to account for different cell culture media compositions and to identify pathways that maximize the formation of particular product metabolites.[101] The validity of FBA has been demonstrated in metabolic engineering studies which have shown it to correctly predict the rate of substrate utilization and cellular growth in bioreactor settings.[102] Other iterations of FBA have been developed such as flux variability analysis (FVA)[103] which explores alternative steady state fluxes that can optimize the objective function equally well. Similarly, parsimonious FBA[104] identifies fluxes involving the fewest active genes while still maintaining optimal biomass growth. Much effort has also been placed to incorporate omics data with FBA to more faithfully capture the metabolic phenotype and improve flux inferences. Åkesson et al. was one of the first groups to do this by using gene expression data to constrain internal fluxes to zero if the associated genes were lowly expressed.[105] This approach led to an explosion of subsequent methods that used gene expression data to either identify active/inactive reactions,[106–108] set lower and upper flux boundaries,[109–111] or correlate transcript levels with metabolic fluxes.[112,113] Protein abundance data has likewise been considered to set FBA flux boundaries[114], constrain fluxes by proteome allocation,[115] and correlate the

enzyme-flux levels.[116]

Kinetic models

Kinetic models of metabolism refer to mathematical models of biochemical reaction rates that account for the enzyme kinetics and concentrations of the participating substrates. In contrast with stoichiometric methods such as FBA, kinetic models take the form of ordinary differential equations (ODEs) which describe the time-dependent dynamics of metabolic reactions. This section on kinetic models is limited to modelling metabolic fluxes since a type of sensitivity analysis is discussed later in Section 2.3.1.

Perhaps the earliest kinetic framework for modelling the interactions between substrates is the Law of Mass Action which was formulated in 1864 by the mathematician Peter Waage and chemist Cato Guldberg.[117] The Law of Mass Action states that the rate of a reaction is proportional to the concentration of the participating reactants. For reversible reactions that are at equilibrium, the product and substrate concentrations remain in a fixed ratio. For a theoretical reaction $a \cdot A \rightarrow b \cdot B + c \cdot C$, the equilibrium constant is

$$K = \frac{[B]^b \cdot [C]^c}{[A]^a}, \tag{2.4}$$

where a, b, and c are the stoichiometric units of species A, B, and C. This reaction is considered an elementary reaction if $a = 1$ since it involves only a single substrate A undergoing a transformation to produce b units of B and c units of C. The flux through this reaction would be modelled as $v = k \cdot [A]$, where k is the mass action kinetic parameter. Mass-action kinetics can also model flux through reversible reactions. If the previous theoretical reaction was reversible with $b = 1$ and $c = 1$, the reverse flux (v_r) would be modelled as the bimolecular reaction $v_r = k_r \cdot [B] \cdot [C]$. Hence, the net flux through this (now reversible) reaction would be $v_{net} = k \cdot [A] - k_r \cdot [B] \cdot [C]$. A positive v_{net} implies that net flux is carried in the forward direction while a negative

value indicates net flux from B and C back to A.

Mass-action equations are often used to describe diffusion-based processes, such as lipid transport across compartments,[118] and are assumed to occur in a well-mixed environment with a fixed volume. Mass-action kinetic models of enzymatic reactions are less common but some effort has been made to parameterize these models for large-scale metabolic networks from metabolite and protein abundance data.[119]

More complex kinetic models have since been formulated which incorporate kinetic properties of enzymes into the equations. Michaelis-Menten kinetics are one of the best known and widely used models which can accurately predict fluxes for a large class of enzymatic reactions.[120] The rate of the reaction is assumed to be proportional to the concentration of the enzyme-substrate complex with parameters K_m (Michaelis constant) reflecting the affinity of the enzyme for its substrate and k_{cat} (catalytic constant) reflecting the maximum rate of a single enzyme active site converting a substrate into a product molecule. While this model is named after Michaelis and Menten who formulated this model in 1913,[121] similar derivations were published a year later by Van Slyke and Cullen.[122] Briggs and Haldane popularized Michaelis-Menten kinetics based on their quasi-steady state approximation that eliminated the need to assume equilibrium or irreversible binding between the substrate and enzyme.[123] The reaction equation describing a Michaelis-Menten reaction is

$$\mathrm{S} + \mathrm{E} \underset{k_r}{\overset{k_f}{\rightleftharpoons}} \mathrm{E\text{-}S} \xrightarrow{k_{cat}} E + P, \tag{2.5}$$

where S is the substrate, P is the product, E is the free enzyme, and E-S is the enzyme-substrate complex. Note that the formation of the enzyme-substrate complex is reversible with a forward and reverse kinetic parameter while the formation of product is irreversible with a single kinetic parameter. By enforcing several assumptions, notably the Briggs-Haldane assumption that the enzyme-substrate binding is fast compared to

complex dissociation or product formation (such that $k_f >> kr$ and $k_{cat} > k_r$), the Michaelis-Menten equation describing the reaction rate is

$$v = \frac{V_{\max} \cdot [S]}{K_M + [S]} \tag{2.6}$$

where

$$K_M = \frac{k_r + k_{cat}}{k_f} \quad \text{and} \quad k_{cat} = \frac{V_{\max}}{[E] + [\text{E-S}]}. \tag{2.7}$$

Michaelis-Menten reactions can also account for various types of inhibition mechanisms that can reduce enzymatic activity and diminish flux through the reaction. These include competitive inhibition, where other substrates bind to the free enzyme; noncompetitive inhibition, where other substrates bind to different sites on the free or complexed enzyme; and uncompetitive inhibition, where other substrates bind to the enzyme-substrate complex.[124]

While other types of kinetic models have been formulated in recent years (e.g. lin-log kinetic[125]), much effort has been devoted to constructing Michaelis-Menten kinetic models for GEMs. A major challenge in this endeavour is the lack of experimentally-derived kinetic parameters for enzymes across different organisms. Some experimental values are contained in enzyme databases such as BRENDA[126] and SABIO-RK,[127] but most enzyme kinetics remain uncharacterized in large-scale metabolic networks. This problem is currently being addressed through several optimization-based methods to fit kinetic parameters that both satisfy experimentally observed flux data and experimentally deduced kinetic parameters.[128–131] Sometimes multiple sets of kinetic parameters may equally describe the experimental data,[132] possibly due to model sloppiness[133] and parameter unidentifiability, which may require further experimental data.[134] Other methods leveraging protein structure databases and protein abundance data have been developed to infer kinetic parameters.[135,136]

Bayesian statistical models of flux

While less common than the methods described above, several attempts have been made to develop statistical models of flux. These models contrast with FBA and kinetic models in that they formalize the uncertainty of identifying steady state fluxes in metabolic networks. As discussed in the previous section on FBA, the solution space of steady state fluxes is vast and poorly constrained by the stoichiometry matrix alone. One may imagine obtaining flux observations from stable isotope tracing studies, but these values may still be explained through many possible sets of steady state fluxes. This uncertainty only increases in larger-scale networks where there is a smaller proportion of observed fluxes and greater ambiguity in the flux solution space.

Bayesian inference is a very natural approach to resolving this problem of flux uncertainty when only a subset of metabolic fluxes are known. Bayesian approaches assume the existence of multiple steady state fluxes that could possibly explain the observed flux values. These possibilities are realized as a joint posterior distribution of steady state fluxes conditioned on some observed flux values.

Bayesian FBA[137] and BayFlux[138] are two notable statistical methods that use Bayesian statistics to estimate a posterior distribution of steady state fluxes conditioned on observed flux data. Bayesian FBA formulates a joint distribution of intracellular or extracellular metabolic fluxes and metabolite changes, assuming multivariate Gaussian priors. The parameters in this model are the mean flux and metabolite accumulation/depletion values and a pairwise covariance matrix which may be simulated or estimated from stable isotope tracing data. After conditioning the model on flux observations and covariances, the posterior distribution is estimated by Gibbs sampling. BayFlux takes a similar approach to Bayesian FBA, but conditions the model on both extracellular fluxes and the isotopic distribution of endogenous metabolites acquired from stable isotope tracing studies. Both methods have been used to estimate steady

state fluxes in GEMs containing over 2,382 metabolites and 1,668 reactions.[137,138]

2.3 Methods to analyze flux data

Once estimated, metabolic fluxes can be analysed through a variety of techniques. All methods described in this section are based on mathematical—rather than statistical—analyses. One possible reason for this is because mathematical models are often formulated when the structure of a problem is well-defined. For metabolic networks, this is especially true since the structural relationship between metabolites and reactions in many metabolic pathways is well understood. Furthermore, statistical techniques rely on analysing many datasets to identify underlying patterns which has historically been infeasible to estimate experimentally or even simulate in large-scale networks. I first begin by briefly reviewing metabolic control analysis (MCA) which is one of the oldest methods to analyse kinetic models of metabolism. I then describe several stoichiometric methods to analyse fluxes, including elementary flux mode (EFM) analysis and other types of stoichiometric pathways.

2.3.1 Metabolic control analysis

MCA is a mathematical framework for understanding the relationships between components in a metabolic flux network under steady state. There is a rich mathematical history behind MCA and this section is kept intentionally short since a more comprehensive review is beyond the scope of this thesis. For clarity, I will also refer to the standardized MCA terminology put forth by Westerhoff et al.[139]

MCA emerged independently in the 1970s from works by Kacser and Burns[27,28] and Heinrich and Rapoport.[29,30] In their seminal papers, both authors addressed the question of how steady state pathway flux is modulated by enzyme and substrate levels

through a series of biochemical reactions. They reasoned that a rigorous mathematical approach to modelling biochemical systems was necessary to quantify the effect of each component on pathway flux. This long-term change in pathway flux due to a perturbation of a variable was called "control," giving rise to the name metabolic control analysis.[140]

A central goal of MCA is to quantify how steady state pathway flux, a global property of a biological system, is controlled by individual variables such as enzyme, substrate, or inhibitor levels. MCA derives several types of control coefficients to denote the control of one variable exerts on another. Flux control coefficients ($C^J_{E_i}$), for example, are defined as the control exerted by a single enzymatic step in pathway J on the overall pathway flux. This is equivalent to the ratio between the relative change in steady state fluxes and levels of a given enzyme i:[27]

$$C^J_{E_i} = \frac{dJ/J}{dE_i/E_i} = \frac{E_i \cdot dJ}{J \cdot dE_i} = \frac{d\ln J}{d\ln E_i} \approx \frac{J\%}{E_i\%} \tag{2.8}$$

For example, a flux control coefficient of 0.15 or $C^J_{E_i} = 0.15$ suggests that increasing the level of enzyme i by 1% would increase the flux along pathway J by 0.15%. A second type of concentration control coefficient ($C^{S_j}_{E_i}$) is similarly defined as the control exerted by changing the level of enzyme i on steady state level of substrate j or

$$C^{S_j}_{E_i} = \frac{dS_j/S_j}{dE_i/E_i} = \frac{E_i \cdot dS_j}{S_j \cdot dE_i} = \frac{d\ln S_j}{d\ln E_i} \approx \frac{S_j\%}{E_i\%}. \tag{2.9}$$

An important discovery of MCA was that the sum of all flux control coefficients is equal to one for a given metabolic pathway, and that the sum of all concentration control coefficients in a metabolic pathway is equal to zero. These findings form the basis of the summation theorem[26,27,29,141] for flux control

$$\sum_{i=1}^{n} C^J_{E_i} = 1 \tag{2.10}$$

and concentration control

$$\sum_{i=1}^{n} C_{E_i}^{S_j} = 0, \tag{2.11}$$

where n is the number of reactions in pathway J.

Another important function of MCA is to understand how enzyme levels change individual reaction rates and how these local properties give rise to systematic properties regarding the entire metabolic pathway. In addition to flux/concentration control coefficients, MCA also defines elasticity coefficients which refer to the change in an isolated reaction rate with respect to the concentration of a substrate, product, enzyme or effector such as an inhibitor. These elasticity coefficients quantify the local control of these modifiers over the flux of an individual reaction when all other modifier concentrations are held constant within the metabolic pathway. The elasticity $\epsilon_{S_j}^{v_i}$ describes the relative change in flux along a single reaction v_i by modifying the concentration of substrate S_j while keeping all other substrate levels constant. Following flux and concentration control coefficients, their equations are defined as

$$\epsilon_{S_j}^{v_i} = \left(\frac{\partial v_i / v_i}{\partial S_j \; S_j}\right) = \left(\frac{S_j}{v_i}\frac{\partial v_i}{\partial S_j}\right) = \left(\frac{\partial \ln v_i}{\partial \ln S_j}\right) \approx \frac{v_i\%}{S_j\%}, \tag{2.12}$$

where the concentrations of other substrate modifiers are held constant. Although elasticities measure local changes to individual reactions, knowledge of these properties can give rise to global properties describing flux through a metabolic pathway. This relationship is described by the connectivity theorem of MCA which relates elasticities to both flux control and concentration control coefficients.[27,139] Applying both summation and connectivity theorems allows one to identify global control coefficients from knowledge of the local elasticity coefficients[29,31,140,141] Since it can be challenging to perturb single reactions within a metabolic pathway, it has also been shown that one can identify the elasticity coefficients in a metabolic pathway from the flux control and metabolite control coefficients.[142]

2.3.2 Elementary flux modes (EFMs)

The EFM is a type of steady state pathway that emerged through the mathematical analysis of stoichiometry matrices and metabolic fluxes under steady state. This stoichiometric pathway analysis method contrasts with MCA in that it operates on the mass-balance constraints imposed by reaction stoichiometries and does not consider the kinetic properties of biochemical reactions. Before formally defining the EFM, it should be noted that these stoichiometric analyses predate MCA by nearly a decade, going as far back to work by Milner.[143] Further stoichiometric analyses were developed in parallel of MCA, including FBA (described in Section 2.2.2),[101] and other analyses of stoichiometric systems under steady state.[144,145] By 1994, EFM analysis was the latest metabolic pathway analysis approach, proposed by S Schuster and Hilgetag, and has since remained an important method for studying metabolic flux networks.

The EFM can be defined as a minimal set of reactions carrying steady state flux within a metabolic network.[37] This minimal property implies that steady state flow within the pathway cannot be maintained by a subset of reactions within an EFM. In other words, an EFM may span source and sink metabolites or encode a cycle within internal metabolites, but that pathway can never contain another EFM within it. The EFM is rather elegant to interpret since it emphasizes how metabolic networks defined by a stoichiometry matrix can be decomposed into functional units of steady state flux.[146]

Understanding how one arrives at the EFM definition requires a mathematical analysis of stoichiometry matrices. Recall from Section 2.2.2 that the mass-balance constraints for a stoichiometry matrix define a system of linear equations that constrain the solution space of steady state fluxes to a flux cone. While there may be an infinite number of steady state fluxes, it is often useful to ask whether those solutions can be described by a combination of smaller steady state metabolic pathways. This question

is a special case of finding the basis for the nullspace in linear algebra.[141] Since basis vectors are not unique, however, a more sophisticated approach is required to explain the observed fluxes in terms of a fixed number of steady state pathways.

It can be shown that EFMs arise from the convex analysis of the flux cone. First, Equation (2.2) is expanded by splitting all reversible reactions in the stoichiometry matrix into pairs of irreversible ones. This leads to

$$S' \cdot v' = 0, \tag{2.13}$$

where S' is the m' by n' stoichiometry matrix of irreversible reactions and v' is the corresponding vector of irreversible steady state fluxes. The solution space of steady state fluxes in Equation (2.13) also forms a flux cone $\mathcal{C}$ defined by

$$\mathcal{C} = \{v' | S' \cdot v' = 0, v' \geq 0\}. \tag{2.14}$$

The edges of this flux cone define the EFMs of the stoichiometry matrix.[37,58] Similar to finding a basis for the nullspace, any point within the volume of the flux cone can be written as a positive linear combination of the flux cone edges. This explains why EFMs are defined as minimal sets of reactions or genetically independent sets of enzymes, since no edge in the flux cone can be written as a combination of other edges. A key difference compared to the nullspace approach to analysing stoichiometry matrices is that there is always a finite number of EFMs characterizing the flux cone.[147]

Each EFM e_k can be written as vector of length n' with elements corresponding to the number of times each reaction occurs throughout that pathway. If e_k are the EFMs of the flux cone for $k = 1, 2, \ldots, E$, the EFM weight decomposition theorem states that any set of steady state fluxes v'' can be constructed from a positive, linear combination

of these pathways or

$$v'' = \sum_{k=1}^{E} w_k \cdot e_k \quad \forall\, w_k \geq 0 \tag{2.15}$$

where w_k is the weight of EFM k.

The EFM has several properties that make it ideal for analysing metabolic fluxes. As discussed in Chapter 1, EFMs are steady state pathways meaning they are a possible explanation for how metabolites travel through or within a metabolic network under the steady state assumption. As minimal pathways, steady state flux is lost if any reaction in an EFM is deleted. Equivalently, EFMs are also nondecomposable pathways where an EFM cannot contain another EFM. These conditions lead to a finite number of EFMs characterizing a given metabolic network, regardless of the steady state fluxes across all reactions. This properties ties into the "elementary" nature of EFMs, as any set of steady state fluxes in a network can be represented as positive, linear combination of their EFMs. The analysis of EFM weights also does not require enzyme kinetic information and can be applied to steady state fluxes estimated from stoichiometric, kinetic, or statistical models of flux. Finally, EFMs are the optimal metabolic pathways for maximizing flux per unit total of protein.[148,149]

Since fundamental problems regarding EFM enumeration and flux decomposition are described in Chapter 1, this remaining section describes several applications of EFM analyses. EFM analyses have been applied to investigate global properties of cellular metabolism such as metabolic robustness. This property can be defined as the ability of a cell to maintain metabolic activity against perturbations.[25] The number of EFMs is a measure of metabolic robustness since it quantifies the number of redundant pathways producing metabolites of interest. Stelling et al., in particular, showed that the biomass growth of *E. coli* was similar between mutants with enzymatic deletions and wildtype strains until a significant number of EFMs were eliminated from the metabolic network.[38] EFMs have also been used to design cell strains for metabolic

engineering purposes. Carlson and Srienc were the first to develop a method to identify the most efficient EFMs in *E. coli* that converted the fewest stoichiometric units of glucose and oxygen into the greatest stoichiometric units of carbon biomass.[41,42] Subsequent studies have applied this same procedure to engineer metabolite production in other bacterial and yeast systems.[150,151] A combination of FBA and EFM analysis has also been used to predict *in silico* gene deletions that are correlated with increased flux through desirable metabolic products.[152,153] Within the realm of human health and disease, fewer attempts have been made to analyse flux data. A notable example is Rezola et al.[154] who formulated a method to identify differentially active EFMs from differentially upregulated genes identified between disease and control conditions. Their approach used a hypergeometric test to identify "characteristic" EFMs involving a greater-than-expected number of highly-expressed genes and as few as possible lowly-expressed genes. Applying their method to bulk gene expression data across different human tissues revealed tissue-specific EFMs which further mapped to well-known metabolic pathways contained in the HumanCyc database.[154]

2.3.3 Other types of stoichiometric pathway analyses

EFMs are not the only type of stoichiometric pathway definition. Since their proposal, EFMs have inspired several related stoichiometric pathways based on the notion of non-decomposability. These pathways are also less numerous than EFMs and therefore computationally faster and more feasible to compute in large-scale networks. That being said, the majority of these pathways have been analysed at the structural level without fluxes, since fewer pathways does not guarantee a unique pathway flux decompositions.

Extreme pathways (EPs)

The extreme pathway (EP) is a type of minimal, steady state pathway proposed by Schilling et al.[147] They are mathematically defined as minimal sets of reactions that form the edges of a flux cone defining the solution space of steady state fluxes for a given stoichiometry matrix.[155] This definition makes EPs rather similar to EFMs. In a stoichiometry matrix consisting exclusively of irreversible exchange and internal reactions, the network EFMs and EPs are identical. It is only when a stoichiometry matrix contains reversible exchange reactions that EPs can be viewed as a subset of EFMs. While EFMs can be defined as genetically independent pathways,[156] EPs are defined as systematically independent pathways since a non-negative linear combination of these pathways can describe all feasible steady state fluxes in a metabolic network.[147] This property also makes EPs more convenient to work with than EFMs since there may be fewer EPs than EFMs and therefore more computationally feasible to compute in large-scale networks.[157,158] That being said, it has been well-documented that there are often many ways to decompose fluxes onto EPs.[159] This computational convenience is also compensated by a potential loss in important steady state pathways describing the structure of the metabolic network. Hence, EP analysis is often inappropriate for assessing metabolic robustness in networks with reversible exchange reactions.[160]

Minimal cut sets (MCSs)

Minimal cut sets (MCSs) were proposed a few years after EPs in 2004 by Klamt and Gilles. The MCS is not a type of pathway *per se* and was originally defined as a minimal set of reactions whose removal leads to the failure of a user-specified objective function (such as the biomass reaction). This definition was later generalized to include the removal of nodes which, while not generally possible in metabolic networks, is possible in cellular networks that include protein receptors and kinases which may be inhibited

by antagonists.[162]

MCSs were historically computed from EFMs and required complete knowledge of all network EFMs to enumerate the corresponding MCSs. These sets of k reactions were useful for studying metabolic robustness by characterizing combinations of lethal reactions required to maintain biomass growth in a cell[163]. More recent algorithmic work has greatly improved MCS enumeration[164] by computing these sets without the need to first enumerate the corresponding network EFMs.[165,166] These advancements are enabling MCS analysis in meta-GEMs modelling the metabolic interactions between multiple bacteria species containing over three thousand reactions.[167]

Elementary conversion modes (ECMs)

Elementary conversion modes (ECMs), proposed by Urbanczik and Wagner, are defined as non-decomposable vectors that map minimal stoichiometric quantities of source and sink metabolites.[95] To understand why these pathways were developed, consider how there may exist multiple EFMs that involve the same stoichiometric number of nutrient sources and product metabolites leaving the network. These EFMs will differ in the configuration of internal reactions explaining those source metabolite movements. For the purpose of understanding input/output metabolite relationships, EFMs involving the same stoichiometric numbers of source and sink metabolites can be considered redundant. ECMs therefore do not consider the distinct sets of internal reactions that explain the relationship between external metabolites. This simplicity means there are orders of magnitude fewer ECMs than EFMs,[168,169] although ECMs analysis cannot be used to investigate the possible mechanisms explaining the movement of metabolites through a metabolic network.

Yield vectors (YVs)

The problem of identifying the subset of biologically relevant EFMs that explain source-to-sink metabolic flows has also been approached through the concept of yield analysis proposed by Song and Ramkrishna.[170] Yield analysis involves recasting EFMs into yield vectors (YVs) which are the stoichiometric ratios between a given pair of substrate and product metabolites. The resulting transformation leads to a subset of YVs whose positive linear combination form the edges of a bounded convex hull describing the space of substrate-to-product conversion yields. This convex hull can be analysed structurally to identify biologically relevant YVs that can cover the majority of the yield space or the most efficient YVs converting substrates into products. This can be achieved through a top-down approach by iteratively removing YVs that negligibly reduce the convex hull volume, or a bottom-up approach by starting with an initial set of three YVs and sequentially adding YVs that maximize the updated volume of the convex hull. When steady state fluxes are observed, the corresponding data can be visualized as a point within the convex hull of the yield space. Song and Ramkrishna identify the subset of YVs that best explain these yields by identifying the largest minimal convex hull (a set of three YVs) that enclose the observed data point. More recent algorithmic advancements have enabled yield analysis without the need to explicitly enumerate all network EFMs.[171]

Elementary carbon modes (ECaMs) and carbon flux paths (CFPs)

A final type of stoichiometric pathway is the elementary carbon mode (ECaM) proposed by Pey et al.[172] The ECaM is defined as an EFM but with respect to a single carbon atom. Like EFMs, these pathways may form simple paths or simple cycles of carbon atom flux. While there may be more ECaMs than EFMs, these pathways are

much simpler to analyse given they are strictly non-branching pathways or cycles. The solution space of carbon atom fluxes also forms a convex polyhedral cone similar to the flux cone which represents the solution space of steady state reaction fluxes.[94,95] This means that ECaMs inherit the EFM property of reconstructing all feasible carbon atom fluxes from a positive, linear combination of ECaMs.

Pey et al. also developed a related stoichiometric pathway called the carbon flux path (CFP) which is a sequence of reactions involving carbon exchange fluxes[173] which was later refined to account for atom mapping data.[174] While similar to ECaMs, CFPs are not necessarily minimal pathways and are enumerated up to a length of k reactions. These pathways have been used to study the connectivity of metabolite graphs in response to gene deletions.[175]

2.4 Simplified molecular input entry system strings

In Chapters 6 and 7 SMILES strings are used in the proposed computational workflow. Briefly, SMILES strings encode the three-dimensional structure of chemical species as plain text.[176] This is very useful for representing chemical compounds in a format that can be recognized by computer software programs for predicting atom mappings within reactions.[177] The remaining section describes how SMILES strings encode molecular structures, referencing selected compounds described in Table 2.1 and Figure 2.1.

SMILES strings represent individual atoms in a molecule using their standard elemental abbreviations. For example, a carbon atom is denoted as a `C` while an oxygen atom is denoted as `O`. The one exception to this rule are hydrogen atoms which can either be explicitly or implicitly represented in a SMILES string. In Table 2.1, the compound methane is denoted by a single `C` with the assumption that it is single bonded to four hydrogens to complete the valence shell. A slightly more complex example is

carbon dioxide which consists of a carbon double bonded to two oxygens denoted with equal signs.

Table 2.1: Examples of SMILES strings encoding selected metabolites.

Metabolite	SMILES string
Methane	`C`
Carbon dioxide	`O=C=O`
Palmitic acid	`CCCCCCCCCCCCCCCC(=O)O`
D-Glucose	`C([C@@H]1[C@H]([C@@H]([C@H](C(O1)O)O)O)O)O`

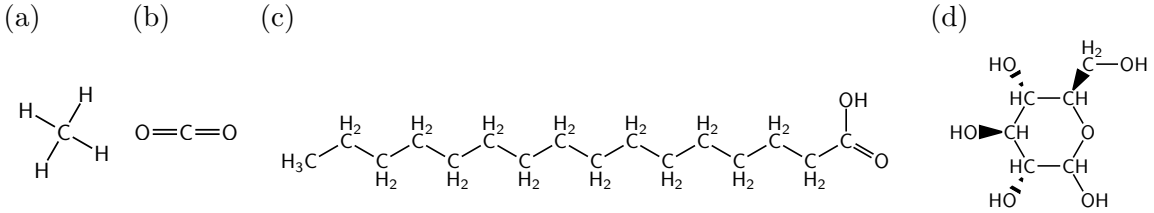


Figure 2.1: Metabolite structures generated from SMILES strings in Table 2.1. (a) Methane. (b) Carbon dioxide. (c) Palmitic acid. (d) D-Glucose.

Connectivity between atoms is denoted by hyphens (`-`) for single bonds, equal signs (`=`) for double bonds, and number signs (`#`) for triple bonds. Since many organic compounds contain aliphatic hydrocarbons, single bonds are not required between carbon atoms that are single bonded to other carbon atoms. An example of this implicit connectivity rule is shown for palmitic acid in Figure 2.1c. The parentheses surrounding the double bonded oxygen indicate this atom (technically an `-OH`) branches off from the preceding carbon to denote part of the carboxyl group.

More complex branching structures are handled by enclosing atoms in nested parentheses. An example of this is shown in the final compound D-glucose. The first carbon is immediately followed by parentheses enclosing the remaining carbon atoms followed by an oxygen. This indicates that the first carbon in the SMILES string corresponds to the 6th carbon of D-glucose located outside of the cyclic ring structure. The square brackets with symbols `[C@H]` and `[C@@H]` represent the chirality of the carbon atom

and its neighbouring hydrogen atom. The `@` sign is read as anticlockwise and `@@` is read as anti-anticlockwise (or simply clockwise). Referencing Figure 2.1, the clockwise order corresponds to the wedged bonds while the anticlockwise order corresponds to the dashed bonds. Lastly, ring structures are explicitly broken into a linear chain with ring closure labels used to denote atoms that are fused together. These label are indexed at 1 and, for the SMILES string encoding D-glucose in Table 2.1, link the 5th carbon and the ring oxygen.

Due to the flexible representation of SMILES strings, there are often numerous ways of writing SMILES strings corresponding to a given molecule. For example, one could reverse the order of a SMILES string or arbitrarily break a ring structure into a linear chain at an arbitrary atom. The process of canonicalization refers to the standardization of SMILES strings through a particular SMILES generation algorithm. *Canonical* SMILES strings should not be confused with *isomeric* SMILES strings which explicitly represent molecular chirality. Both canonical and isomeric SMILES strings are provided in compound databases such as PubChem.[178]

2.5 Markov chains

Markov chains are a type of probabilistic model known as a stochastic (or random) process which describe the sequence of possible events occurring across time. They are grounded under the fundamental assumption that the probability of some future event is only dependent on the present state. This is known as the Markov property (or Markov assumption/constraint), and more generally referred to as "memorylessness" in probability theory.[179]

To make this definition clearer, let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of random variables describing the state of some dynamical system. The possible values of each X_n are referred to as the state space $S = \{x_1, x_2, \ldots, x_k\}$. In other words, this stochastic

process describes the sequence of outcomes associated with random variable X over time n. This stochastic process is called a Markov chain if

$$\Pr(X_{n+1} = x_{n+1} | X_n = x_n, \ldots, X_0 = x_0) = \Pr(X_{n+1} = x_{n+1} | X_n = x_n) \tag{2.16}$$

for all $n \in N$ and $x_1, x_2, \ldots, x_{n+1} \in S$. This Markov chain is specifically referred to as a discrete-time Markov chain (DTMC) since the evolution of X_n occurs over discrete time steps n. Note how under the Markov property, the probability of future event at time $n+1$ is conditioned only on the probability of the current event at time n. It is often assumed that the probabilities of events do not depend on the n, so $\Pr(X_{n+1} = j | X_n = i) = P_{ij}(n) = P_{ij}$. Markov chains satisfying this property are called time-homogeneous where P_{ij} refers to the transition probability of moving from state i to state j. For Markov chains with finite state spaces, the transition probabilities of moving from state i to j can be represented as stochastic matrix P.

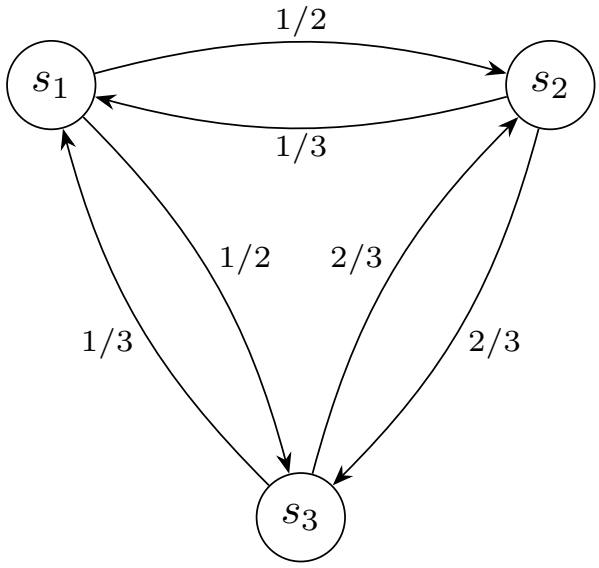


Figure 2.2: An example time-homogeneous DTMC.

Consider the following Markov chain in Figure 2.2. The state space S consists of three states s_1, s_2, and s_3. Since this Markov chain is time-homogeneous, the probabilities of moving between states is independent of the discrete time steps. The probabilities

for moving from each state i to j can be written in matrix form as

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix} = \begin{bmatrix} 0 & 1/2 & 1/2 \\ 1/3 & 0 & 2/3 \\ 1/3 & 2/3 & 0 \end{bmatrix}. \tag{2.17}$$

For example, P_{12} denotes the transition probability of moving from state s_1 to state s_2. The transition probability matrix also allows states to transition to themselves although no such transitions occur in this example so $P_{ii} = 0$. Matrix P is known as a right stochastic matrix since $\sum_{j \in S} P_{ij} = 1$ for every row i. It is possible for a Markov state to have no outgoing transitions in which case $P_{ii} = 1$. These are known as absorbing states since once this state has occurred, it is impossible to leave that state.

Markov states can also be classified according to other properties that describe the interactions between states. For example, states i and j are said to communicate with each other if there exists a path (referring to a sequence of transitions) from state i to j and from j back to i. States that communicate with each other form communication classes and a Markov chain is said to be irreducible when all states can communicate with each other.

Markov chains are commonly studied to understand the dynamic evolution of events occurring over time. Starting at a given state i, these dynamics can be simulated by repeatedly transitioning across states based on the transition probability matrix. Returning to Figure 2.2, one may ask what is the probability of transitioning from s_1 to s_2 in two steps. These probabilities are readily computed by multiplying matrix P

by itself yielding

$$P^2 = P \cdot P = \begin{bmatrix} 0 & 1/2 & 1/2 \\ 1/3 & 0 & 2/3 \\ 1/3 & 2/3 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1/2 & 1/2 \\ 1/3 & 0 & 2/3 \\ 1/3 & 2/3 & 0 \end{bmatrix} = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 2/9 & 11/18 & 1/6 \\ 2/9 & 1/3 & 11/18 \end{bmatrix}. \tag{2.18}$$

For N steps, the N-step transition probabilities are equal to P^N. Focusing on state s_1, the two-step probability of returning back to s_1 is $1/3$ since there is a $1/2 \times 1/3$ probability that s_1 transitions to either state s_2 or s_3 before returning back to s_1. As N approaches infinity, one may ask whether the probability of occupying a given state in S converges onto a long-term probability distribution. These long-term probabilities are notated as π and are also referred to as the stationary distribution or steady state probabilities of the Markov chain. These probabilities are unique in Markov chains with a finite state space that are irreducible and aperiodic. The term aperiodic means that there is no certainty regarding the number of transitions required for state i to eventually revisit itself. Conversely, a Markov chain is periodic when states can only be revisited at regular intervals. Markov chains that are both irreducible and aperiodic are considered ergodic where the limit

$$\pi_j := \lim_{n \to \infty} P_{ij}^n \tag{2.19}$$

exists for every state j, meaning the long-term behaviour of the Markov chain does not depend on the initial state i. All steady state probabilities satisfy

$$\pi_j = \sum_{i \in S} \pi_i \cdot P_{ij} \quad \text{where} \quad \sum_{i \in S} \pi_i = 1 \tag{2.20}$$

and can be computed in several different ways. A common "brute force" method is to repeatedly exponentiate matrix P until the probabilities in each row remain un-

changed (indicating convergence towards the stationary distribution). The steady state probabilities can also be computed more explicitly by noticing that Equation (2.20) represents a system of linear equations. Constraining all steady state probabilities to sum to one, the stationary distribution can also be solved algebraically or numerically as a system of linear equations. Revisiting Figure 2.2, these steady state probabilities are

$$\pi = [1/4 \;\; 3/8 \;\; 3/8] \tag{2.21}$$

which one can confirm remain unchanged when multiplied by P since

$$\pi \cdot P = [1/4 \;\; 3/8 \;\; 3/8] \cdot \begin{bmatrix} 0 & 1/2 & 1/2 \\ 1/3 & 0 & 2/3 \\ 1/3 & 2/3 & 0 \end{bmatrix} \tag{2.22}$$

$$= \begin{bmatrix} 1/4 \cdot 0 + 3/8 \cdot 1/3 + 3/8 \cdot 1/3 \\ 1/4 \cdot 1/2 + 3/8 \cdot 0 + 3/8 \cdot 2/3 \\ 1/4 \cdot 1/2 + 3/8 \cdot 2/3 + 3/8 \cdot 0 \end{bmatrix}^{\intercal} = \begin{bmatrix} 6/24 & 9/24 & 9/24 \end{bmatrix} \tag{2.23}$$

$$\pi \cdot P = [1/4 \;\; 3/8 \;\; 3/8] = \pi \tag{2.24}$$

While this section has focused exclusively on describing DTMCs, it would be remiss to not mention (time-homogeneous) continuous-time Markov chains (CTMCs). Unlike a DTMC, a CTMC models the time t between a sequences of n events. A continuous-time stochastic process is called a CTMC if

$$\Pr(X(t_{n+1}) = j | X(t_n) = i) \tag{2.25}$$

for $j, i \in S$ and $t_{n+1} > t_n \leq 0$. Following time homogeneity, Equation (2.5) simplifies to

$$P_{ij}(t) \tag{2.26}$$

since the probability of moving to state j at time t_{n+1} only depends on the current state at time t_n. Since t is continuous rather than discrete, the (one-step) transition probability matrix P used in the DTMC is not applicable as time can vary across a continuum of values. To enforce the Markov property, the holding times should only depend on the current time since transitioning to state i and not the prior time involving previous state sequences leading to $X(t) = i$. This memorylessness property is a feature of both the geometric and exponential distribution. Because only the exponential distribution models continuous random variables, the holding times between CTMC transitions are naturally modelled as exponential random variables (unlike the discrete geometric distribution). Specifically, the time spent in state i is exponentially distributed with rate $Q_{ii} = -\sum_{j \neq i} Q_{ij}$, where $Q_{ij} \geq 0$ is the rate of transitioning from state i to state j.

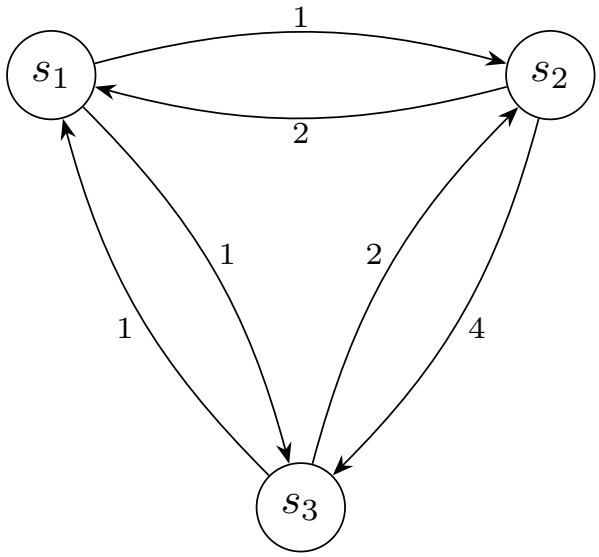


Figure 2.3: An example time-homogeneous CTMC.

To make this explanation more concrete, consider the following CTMC in Figure 2.3. The structure of the Markov chain is identical to Figure 2.2 but is clearly a CTMC since the outgoing arcs are transition rates that need not sum to one unlike transition probabilities. The corresponding transition rates are represented in a matrix referred

to as the generator matrix

$$Q = \begin{bmatrix} -2 & 1 & 1 \\ 2 & -6 & 4 \\ 1 & 2 & -3 \end{bmatrix}. \tag{2.27}$$

Note how the rows of Q sum to zero since the rate of time spent in state i is balanced by the sum of outgoing transition rates to states $j \neq i$.

Every CTMC has a corresponding embedded DTMC which only models the sequence of state transitions without considering the holding times. The transition probabilities of the embedded DTMC can be computed from the generator matrix by setting

$$P_{ij} = \begin{cases} -Q_{ij}/Q_{ii}, & \text{for } j \neq i. \\ 0, & \text{for } j = i. \end{cases} \tag{2.28}$$

Figure 2.4 shows a sample trajectory for the example CTMC. Starting in state s_1 at time $t = 0$, the process holds for τ_1 units of time before instantaneously transitioning to state 2 with probability $P_{12} = \frac{1}{2}$. This process is shown to repeat three more times with a state sequence of s_1, s_2, s_3, s_2, s_3.

Lastly, one can also compute the stationary distribution for a time-homogeneous CTMC similarly to that of a DTMC. These probabilities are notated as q to distinguish them from steady state probabilities π corresponding to a DTMC. For an irreducible and aperiodic CTMC, the steady state probabilities exist when the limit

$$q_j := \lim_{t \to \infty} Pr(X(t) = j | X(0) = i), \tag{2.29}$$

exists for every state j. These probabilities are interpreted as the long-term proportion of time the stochastic process spends in state j, regardless of the initial starting state.

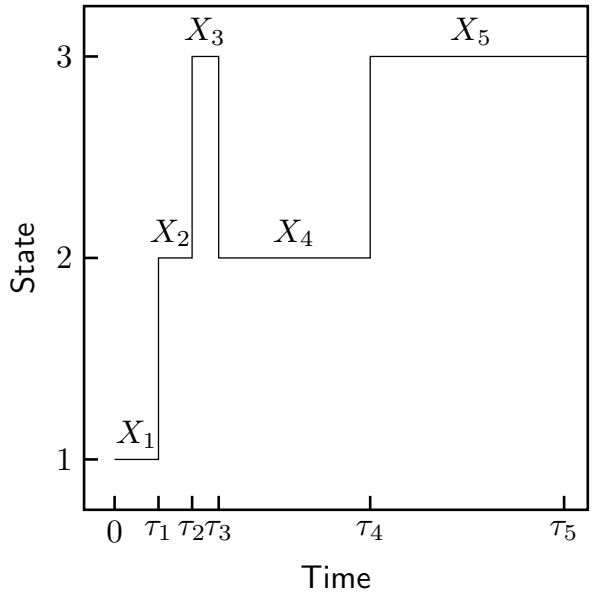


Figure 2.4: Sample trajectory for the CTMC in Figure 2.3 with holding time $\tau_{n+1} - \tau_n$ between state transitions n and $n+1$ follows an exponential distribution with rate $Q_i = -Q_{ii}$. In the long-term, the sequence of states X_n follows an embedded DTMC matching the transition probabilities matching Figure 2.2.

It can be shown that these steady state probabilities satisfy

$$0 = \sum_{i \in S} q_i \cdot Q_{ij} \quad \text{where} \quad \sum_{i \in S} q_i = 1, \tag{2.30}$$

Following the DTMC, the stationary distribution can be computed by solving the system of equations with the constraint that all probabilities sums to one.

In general, CTMCs are used more often in biological applications to model the real-world phenomena where events occur over continuous time intervals. Well-known examples include modelling the concentration of chemical species in a well-mixed environment[180] and protein conformational changes.[181] While both types of Markov chains are discussed in this thesis, DTMCs are covered throughout Chapters 3–7 to model the movement of individual particles and atoms within a metabolic flux network under steady state. Both types of Markov chains are discussed in Chapter 4 where I show how a certain DTMC proposed in this thesis can be reframed as a CTMC in the context of mass-action kinetic models.

3

A Markovian decomposition to uniquely assign elementary flux mode weights in unimolecular flux networks

This chapter includes contents from "A Markov constraint to uniquely identify elementary flux mode weights in unimolecular metabolic networks" by Justin G. Chitpin and Theodore J. Perkins.[182] As an Elsevier journal author, no written permission from Elsevier was required to include this article (licensed under a Creative Commons CC-BY license) in my thesis subject to proper acknowledgement. Minor textual and visual modifications were made to better incorporate the work into this thesis. The algorithm implementation is available at `https://github.com/jchitpin/MarkovWeightedEFMs.jl`, while all data and source codes to reproduce all figures are available at `https://github.com/jchitpin/reproduce-efm-paper-2023`.

3.1 Author contributions

Justin G. Chitpin developed the methods, conducted the experiments, wrote the software, analyzed the results, and wrote the manuscript with guidance from Theodore J. Perkins.

3.2 Abstract

Elementary flux modes (EFMs) are minimal, steady state pathways characterizing a flux network. Fundamentally, all steady state fluxes in a network are decomposable into a linear combination of EFMs. While there is typically no unique set of EFM weights that reconstructs these fluxes, several optimization-based methods have been proposed to constrain the solution space by enforcing some notion of parsimony. However, it has long been recognized that optimization-based approaches may fail to uniquely identify EFM weights and return different feasible solutions across objective functions and solvers. Here I show that, for flux networks only involving single molecule transformations, these problems can be avoided by imposing a Markovian constraint on EFM weights. This Markovian constraint guarantees a unique solution to the flux decomposition problem, and that solution is arguably more biophysically plausible than other solutions. I describe an algorithm for computing Markovian EFM weights via steady state analysis of a certain discrete-time Markov chain, based on the flux network, which I call the cycle-history Markov chain (CHMC). I demonstrate my method with a differential analysis of EFM activity in a lipid metabolic network comparing healthy and Alzheimer's disease patients. This method is the first to uniquely decompose steady state fluxes into EFM weights for any unimolecular metabolic network.

3.3 Introduction

Cellular metabolism consists of carefully-tuned biochemical reactions operating collectively to maintain metabolic homeostasis. These reactions form highly connected metabolic networks whose dynamics regulate bioenergetic processes[183,184] and cell-specific functions.[185] Data from large-scale genomics projects has led to genome-wide reconstructions of bacterial, yeast, plant, and human metabolic networks.[186–189]

When metabolic reaction rates are known, the relationships between metabolites can be modelled as a flux network with metabolites represented as nodes and the fluxes between them as weighted, directed edges. The fluxes may be estimated experimentally by stable isotope tracing analysis and metabolic flux analysis[87,90] or computationally by flux balance analysis,[190] kinetic models,[118,132] and other statistical techniques.[137,191] Computational methods generally estimate fluxes under metabolic steady state, where the production of each metabolite is balanced by its consumption. These flux networks are often constructed to visualize metabolic activity, or analyzed to optimize metabolite production[192] and identify metabolic alterations associated with diseases.[193,194]

Elementary flux modes (EFMs) are a pathway-based approach to study properties of flux networks. The EFM is defined as a non-decomposable, steady state pathway in a flux network.[37] EFMs may either traverse the network, connecting input and output fluxes of external metabolites, or encode looped pathways, such as futile cycles, to maintain steady state. The non-decomposable property specifies that pathway flux is lost if any reaction in an EFM is abolished. Thus, EFMs have been described as functional units of flux expressing any steady state metabolic phenotype.[146]

EFM analyses have been extensively applied in literature to characterize metabolic activity. For example, the number of EFMs producing a given metabolite is an indicator of network robustness.[38] Considerable work has been devoted to quantifying the contribution of individual EFMs towards external metabolite production of bioreac-

tors[41] and correlating EFM activity with metabolic phenotypes.[195] More recently, EFM analyses have predicted metabolic interactions in microbial communities,[196] estimated cellular processes explaining exponential cell growth,[197] and optimized renewable fuel production in bioreactors.[198]

A major computational challenge associated with EFMs is their enumeration in large-scale networks. EFM analyses are typically limited to small biological networks (< 100 reactions) because the number of EFMs explodes combinatorially with network size and number of external metabolites.[64,199] Considerable attention has been devoted to scaling EFM enumeration towards genome-wide networks.[60,160,200] Most recently, advances in parallelized EFM computation have enabled EFM enumeration in medium-sized networks containing as many as 296 reactions with over twelve billion EFMs.[61]

Another key computational problem, and the focus of the current chapter, is to determine how much flux is travelling along each EFM, given the steady state flux along each reaction in a metabolic network. This flux decomposition problem commonly arises in differential EFM analyses to identify metabolic subpathways altered across disease-versus-control or growth conditions.[52,55] However, decomposing steady state fluxes onto EFMs is generally not unique. In a unimolecular flux network with M metabolites, the steady state phenotype can be described by the N reaction fluxes, of which there are at most M^2. In general, the number of EFMs, K, can be as large as M factorial, depending on the structure of the network.

Most existing strategies for solving the flux decomposition problem are dependent on some notion of parsimony and computational elegance, rather than strong biophysical motivation. These optimization-based methods use linear programming (LP), quadratic programming (QP), or mixed-integer linear programming (MILP) to assign EFM weights according to a user-specified objective function. Common optimization-based methods minimize or maximize activity through the shortest or fewest EFM to reconstruct the observed fluxes. However, these methods may return one among many

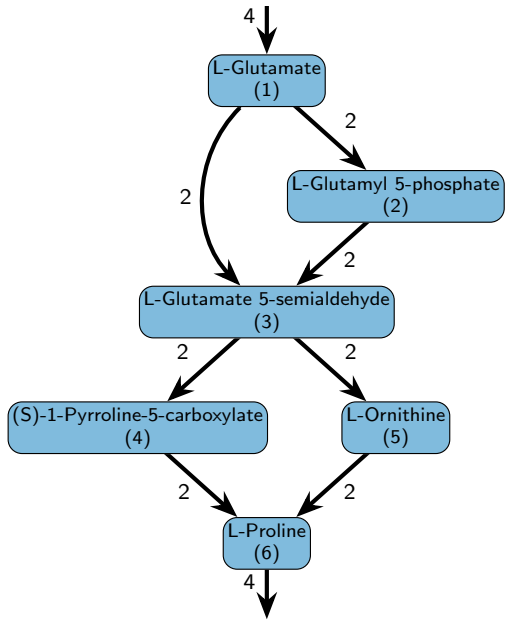
equally good solutions, often without acknowledging the other possibilities.

This issue of non-uniqueness is well-documented in the related metabolic field of flux balance analysis (FBA) which uses similar optimization-based techniques to predict steady state flux distributions.[36,201] Many FBA objective functions have been proposed to quantify the contribution of each reaction towards the metabolic phenotype, producing different flux distributions[202] or a range of equally feasible flux distributions.[103,159,203] Advancements in FBA are solving the non-uniqueness problem by refining the objective function with additional "-omics" data to better describe the metabolic phenotype.[204–207] While this strategy may improve FBA flux predictions, it is unclear how transcript, metabolite, or protein levels could be leveraged to improve EFM weight identification.

Figure 3.1 highlights the flux decomposition problem on two subnetworks taken from the KEGG database.[75] In the first network, glutamate is catabolized via two parallel pathways resulting in proline synthesis. The second network shows tetrahydrofolate (THF) remodelling into three derivatives within the one carbon folate cycle. Figure 3.1a shows five arbitrary sets of EFM weights that perfectly reconstruct the steady state fluxes. Applying an objective function to minimize the number of active EFMs would yield two equally good integer solutions, while a maximization function would return an infinite number of EFM weight solutions. Similar problems arise in Figure 3.1b, where different objective functions may either fail to constrain the solution space, or disagree on the unique set of EFM weights based on the chosen constraints. For example, prioritizing EFM activity through the shortest EFMs in Figure 3.1b would yield the first example solution, while optimizing for the fewest number or activity through the longest possible EFMs would return the second solution.

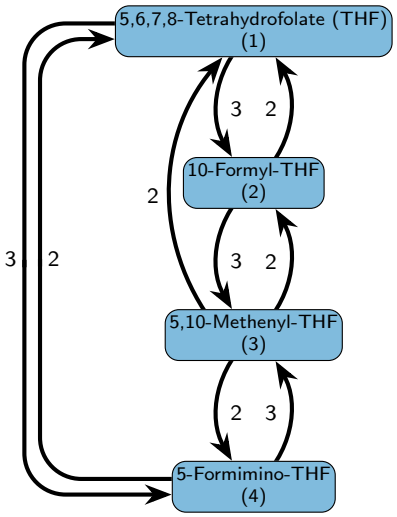
When flux decompositions are not unique, the resulting EFM weights lose their metabolic predictivity. Fundamentally, the chosen EFM weights are a biophysical explanation of how substrates are metabolized in a network. Only a single set of EFM

(a)



EFMs	Example weights				
1. → 1 → 2 → 3 → 4 → 6 →	2	0	1	1.1	1.5
2. → 1 → 2 → 3 → 5 → 6 →	0	2	1	0.9	0.5
3. → 1 → 3 → 4 → 6 →	0	2	1	0.9	0.5
4. → 1 → 3 → 5 → 6 →	2	0	1	1.1	1.5

(b)



EFMs	Example weights				
1. 1 → 2 → 1	2	0	1	0.6	0.5
2. 2 → 3 → 2	2	0	1	0.6	0.5
3. 3 → 4 → 3	2	0	1	0.9	1.0
4. 1 → 4 → 1	2	0	1	0.9	1.0
5. 1 → 2 → 3 → 1	1	1	1	1.3	1.5
6. 1 → 4 → 3 → 1	1	1	1	0.7	0.5
7. 1 → 2 → 3 → 4 → 1	0	2	1	1.1	1.0
8. 1 → 4 → 3 → 2 → 1	0	2	1	1.4	1.5

Figure 3.1: KEGG networks demonstrating EFM weight ambiguity. Nodes and edges denote metabolites and fluxes, respectively. EFMs weights that reconstruct the observed fluxes are shown for the highlighted nodes/edges. (a) Pathway of L-Glutamate to L-Proline (KEGG rn00330). (b) Subnetwork of one carbon folate cycle from (KEGG map00670).

weights should correctly describe the metabolic processes governing the observed, steady state fluxes. One may verify this proposition by considering the set of EFM weights reconstructing the fluxes in Figure 3.1a. Decomposing the fluxes onto EFMs 1 and 4 would suggest glutamate is metabolized separately through two parallel paths that converge on proline. Deleting the reaction between L-glutamate and L-glutamate 5-semialdehyde should therefore abolish flux through (S)-1-pyrroline-5-carboxylate. Conversely, if the fluxes were instead explained by EFMs 2 and 3, I would expect a loss of flux through L-ornithine. This same problem occurs in Figure 3.1b where each chosen set of EFM

weights offers a different biological explanation of THF remodelling.

Despite the well-established use of mathematical optimization, uniquely identifying EFM weights remains a longstanding problem limiting EFM applications. In this chapter, I show how common objective functions described in literature may fail to unambiguously decompose fluxes onto EFM weights. I further argue these methods can produce biophysically unreasonable solutions. A metabolite has no memory of the reactions leading to its current state nor the sequence of reactions it will undergo to complete an EFM in the future. I argue that a more natural approach to assigning EFM weights is to model how a single particle converts between metabolites according to the observed, steady state fluxes. This Markovian assumption captures the intuition of the uniform EFM weight solution best describing the steady state fluxes in Figure 3.1a, and motivates my discrete-time Markov chain model of a single particle converting between metabolites in a unimolecular flux network. By decomposing the Markov chain trajectories into EFMs, I compute a unique probability of a particle transiting an EFM that is proportional to the EFM weights reconstructing those observed fluxes. I solve the problem of computing the EFM transit probabilities via steady state analysis of a new construction I propose, a cycle-history Markov chain. I demonstrate this method by computing EFM weights for the example flux networks in Figure 3.1 and a kinetic model of sphingolipid metabolism of healthy versus Alzheimer's disease patients.

3.4 Methods

3.4.1 Flux networks and elementary flux modes

A unimolecular, steady-state flux network can be represented as a weighted graph $G = (V, E, f)$, where V is the set of nodes corresponding to distinct metabolites, E is the set of directed edges corresponding to chemical transformations of one metabolite

into another, and f are positive real-valued weights of those edges that denote the amount of metabolite per unit time being transformed.[208] This network may be closed or open such that external fluxes enter and leave via source and sink metabolites. In an open network, I assume there is a special node $v_{ext} \in V$ (or possibly multiple such nodes) that represents the external environment. Flux into the network happens along edges $e = (v_{ext}, i) \in E$ for which $i \in V - \{v_{ext}\}$ and $f(e) > 0$. Such nodes i are called sources. Flux out of the network happens along edges $e = (i, v_{ext}) \in E$ for which $v \in V - \{v_{ext}\}$ and $f(e) > 0$, with such nodes i being called sinks. Note that open networks with sources and sinks can be modelled in other ways. However, this way is most notationally convenient for my purposes, because it allows open and closed networks to be treated simultaneously.

I restrict attention to flux networks at *steady state*, meaning that the total flux into each node equals the total flux out: for every $i \in V$, $\sum_{j \neq i} f_{ji} = \sum_{j \neq i} f_{ij}$. I also restrict attention to *strongly-connected* flux networks, where there must be a path in the network from any node i to any other node j. For open networks, this includes the possibility of paths through the external node(s) v_{ext}. If one wants to analyze a flux network with multiple strongly connected components, then each component can be analyzed independently by my proposed method.

The flux network G can be decomposed into a finite set of K EFMs transiting a series of edges in E.[58] In the present context, the EFMs are the set of all simple cycles in G. (For open networks, some authors define the set of EFMs as the union of the simple cycles and the simple paths from any source node to any sink (e.g. Ruckerbauer et al.[209]); however, with my external pseudometabolite(s) v_{ext}, the source-to-sink paths are equivalently represented by simple cycles leaving and returning back to v_{ext}.) The steady state fluxes in G can always be explained as a positive, linear combination of EFMs.[210] Specifically, the weighted sum of all EFMs contributing towards the same reaction should equal the observed, steady state flux of that reaction. This flux

decomposition problem can be solved by identifying a set of EFM weights w satisfying $Aw = f$, where A is an $|E|$ by K matrix with one or zero elements denoting whether a unimolecular reaction is active or inactive in each EFM. When the number of EFMs exceeds the number of linearly independent fluxes, w is undetermined. While optimization-based methods are not guaranteed to constrain the solution space of w, the method I present uniquely identifies w.

3.4.2 Markovian solution to the flux decomposition problem

I propose *Markovian EFM weights* as a way of resolving the issue that finding EFM weights to explain a given set of fluxes is generally an underdetermined problem. I develop my approach in three main steps. First, I define a Markov chain that models a "particle" of metabolite flowing randomly around the flux network and the steady state frequencies with which that particle passes around different simple cycles (i.e. EFMs). Second, I propose a novel construction, the *cycle-history Markov chain* (CHMC), to efficiently calculate those frequencies. Third and finally, I describe how to scale those frequencies to calculate a unique Markovian solution to the EFM weight problem. While I provide intuitive justification of these steps, Appendix A.1 contains a complete proof of the algorithm correctness.

Single particle, discrete-time Markov chain model

My approach to uniquely identifying EFM weights depends on imagining a single "particle" transiting the flux network–repeatedly transformed from one metabolite to another. I make a Markovian assumption that when the particle is in metabolite state $i \in V$, it transitions to a new metabolite state j with probabilities proportional to the outgoing fluxes f_{ij}. This assumption is very natural from a biophysical standpoint. If metabo-

lite molecules are well mixed in a reaction volume along with transforming enzymes, then the future behaviour of any molecule should depend only on its current state. This situation is modelled by a discrete-time, time-homogeneous Markov chain where the set of possible states is V and the transition probabilities are $T_{ij} = f_{ij}/\sum_{j'} f_{ij'}$. Starting from any initial state s_0, the particle will transit a random sequence of states $s_0, s_1, s_2, \ldots$ Because I have assumed the flux network is strongly connected, and by basic theory of Markov chains, with probability 1, that sequence will include all states infinitely many times. Therefore, every time the particle visits a state i, it must eventually return to state i. The sequence of states in between, $s_t, \ldots, s_{t'}$ constitutes a cycle. That need not be a simple cycle–it may comprise multiple nested or interleaved simple cycles. However, it can always be reduced to a unique simple cycle by repeatedly removing the earliest simple cycle found within the sequence—I call this the "first closure reduction rule." For example, a Markov chain state sequence in Figure 3.1b may be 1→4→3→2→3→4→3→1. Metabolite/state 3 is the first to occur twice in the sequence, so I replace the subsequence 3→2→3 with just a single 3, resulting in the reduced sequence: 1→4→3→4→3→1. Now 4 is the first to occur twice, so I replace 4→3→4 with a single 4, resulting in: 1→4→3→1. This is the unique simple cycle to which the original sequence reduces under the first closure reduction rule, and corresponds to EFM 6. In similar fashion, every revisit to any state can be said to have occurred via a particular simple cycle, which of course corresponds to an EFM. Therefore, I propose finding unique EFM weights by setting them proportional to the (steady state) probabilities of the Markovian particle transiting the corresponding simple cycles.

Cycle-history Markov chain

Equating EFM weights with cycle probabilities in the Markov chain in the previous section creates the problem of computing those cycle probabilities which has not been previously addressed in the Markov chain literature. I take the first step in that direction

by defining a CHMC and show how to use it to compute the desired probabilities. The CHMC is related to the single particle Markov chain, but its notion of state is expanded to include enough of the history of the process to determine each time the particle finishes transiting a simple cycle. The CHMC is constructed as follows. I first choose an arbitrary metabolite s_0 in the flux network. (Because my interest is in steady state properties of the Markov chain, the choice of s_0 is not important to the final result, although it may result in some differences in the size of the CHMC being constructed). Each state S in the CHMC corresponds to an entire simple path s_0, s_1, ..., s_m of the Markov chain envisaged in the single particle Markov chain. If there is a reaction transforming metabolite s_m to s_{m+1}, then there is a corresponding transition in the CHMC, but it takes one of two forms. If s_{m+1} is not among s_0, s_1, ..., s_m, then it is simply a transition to the longer simple path–i.e. from the state $S \equiv s_0, s_1, \ldots, s_m$ to a state $S' \equiv s_0, s_1, \ldots, s_m, s_{m+1}$. However, if s_{m+1} is among s_0, s_1, ..., s_m, say $s_{m+1} = s_i$ for $0 \leq i < m$, then this corresponds to closing a simple cycle, and there is a transition from $S \equiv s_0, s_1, \ldots, s_m$ back to the shorter simple path $S' \equiv s_0, s_1, \ldots, s_i$. In either case, the probability of that transition in the CHMC is the same as the transition in the Markov chain: $\mathcal{T}_{S,S'} = T_{s_m, s_{m+1}}$

To make that abstract discussion more concrete, let us first consider the flux network from Figure 3.1a. Suppose the CHMC for that flux network is initialized at metabolite 1. The chain is "grown" recursively by adding children metabolites 2 and 3, and so forth, until there are four simple paths ending at the special external environment node. The resulting CHMC is shown in Figure 3.2a with four additional upstream edges transitioning back to metabolite 1 that mark the completion of a simple cycle for each EFM starting and ending at the root node.

Figure 3.2b shows the CHMC for the second example network. Every EFM appears as a loop somewhere in the transformed network. EFMs involving the initial metabolite occur as precisely one path from root to leaf and then back up to root, such as EFMs

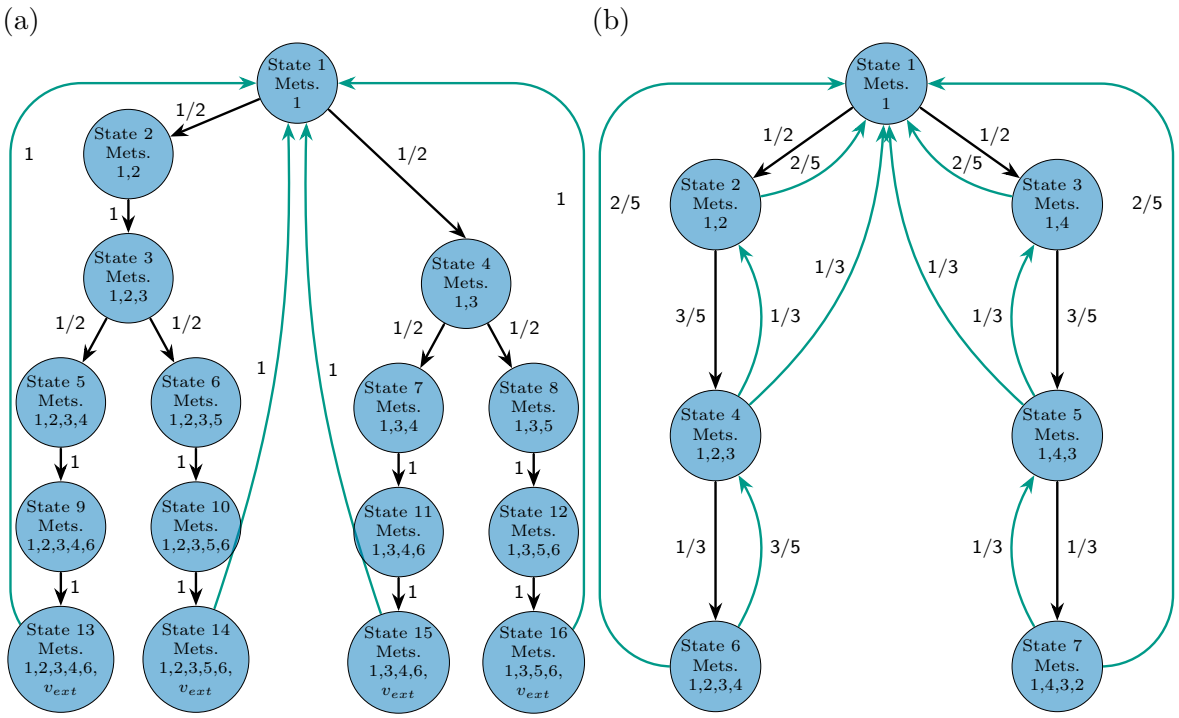


Figure 3.2: CHMCs of networks from Figure 3.1 that are arbitrarily rooted on metabolite 1. Nodes and edges denote metabolite sequences and CHMC transition probabilities, respectively. Coloured arrows represent the completion of a simple cycle corresponding to an EFM.

7 and 8. EFMs not involving the root metabolite also occur somewhere in the CHMC, such as EFM 1 which cycles between metabolites 2 and 1. However, other EFMs may occur in multiple parts of the tree. For instance, EFM 2, which is a loop between metabolites 2 and 3, is represented by two cycles in the CHMC: from state 2 to 4 and back to 2, and also from state 5 to 7 and back to 5.

Computing steady state EFM probabilities

As I have assumed the flux network is strongly connected, the corresponding CHMC will also be strongly connected. This implies the existence of a unique steady state distribution π which is the long-term probability of the particle occupying each CHMC state. To compute the steady state probability of EFM k, I identify all transitions

in the CHMC that correspond to closing that cycle, E_k. For example, returning to Figure 3.2b and EFM 2, I have $E_2 = \{(4,2),(7,5)\}$, as those two edges close cycles involving the 2-3-2 metabolic loop. For other EFMs, E_k may contain just a single transition. I define the steady state probability of transiting EFM k as the sum of the steady state probabilities of those transitions occurring, or:

$$P_k = \sum_{(S,S')\in E_k} \pi_S \cdot \mathcal{T}_{S,S'}. \tag{3.1}$$

In this proposal, the steady state EFM probabilities P_k are proportional to the desired EFM weights w_k and must be scaled appropriately to account for the absolute magnitude of the fluxes. If $f_{total} = \sum_{i,j} f_{i,j}$ is the total flux in the network summed across reactions, and $f_{efm} = \sum_k w_k L_k$ is the total flux obtained by summing EFM weights w_k and their lengths L_k, then the proportionality constant α, where $w_k = \alpha P_k$, should satisfy $f_{total} = f_{efm} = \alpha \sum_k P_k L_k$, so $\alpha = f_{total} / \sum_k P_k L_k$. This scaling ensures not only that the total flux is correct, but that every individual flux in the network is correctly accounted for.

3.4.3 Comparison to optimization-based methods

I benchmarked the Markov model against five optimization-based methods reported in literature. The LP, QP, and MILP objective functions are described in Table 3.1. Briefly, these approaches optimize either the zero-, one- or two-norm of the EFM weights, possibly multiplied by the EFM lengths, and subject to constraints that ensure the fluxes are correctly reconstructed. EFM enumeration and optimization programs were implemented in Julia (version 1.6.5) using the commercially-licensed Gurobi solver (version 9.1.2) and open-source solvers supported in the Julia JuMP package. In total, EFMs weights for the optimization-based methods were computed using eight LP, two QP, and one MILP solver (see Appendix A.2).

Table 3.1: Benchmarked optimization-based approaches.

Description and code	Objective function	Ref.
Minimize L2 norm (L2)	$\min \sum w_k^2$	[55]
Maximize shortest pathway activity (qSPA)	$\max \sum (L_k w_k)^2$	[53]
Maximize shortest pathway activity (lSPA)	$\max \sum (L_k w_k)$	[54]
Minimize shortest pathway activity (lSPA)	$\min \sum (L_k w_k)$	[56]
Minimize active pathways (milAP)	$\min \sum \beta_k = \begin{cases} 1, & \text{if } w_k > 0 \\ 0 & \text{otherwise.} \end{cases}$	[52]

3.5 Results

3.5.1 Markov model uniquely estimates EFM weights in example networks

I first compared the Markov and optimization-based methods on the example networks in Figure 3.1. In the first example network, the steady state probabilities of the particle occupying external metabolite states v_{ext}, denoted by the CHMC nodes 13-16 in Figure 3.2a, were 4.545% because the path probabilities along each prefix were identical. As the probability of completing each simple cycle was 100% from any of those states, the steady state EFM probabilities were 4.545%. Scaling the EFM probabilities to weights yielded the unique solution $w = [1111]$. In contrast, the solutions for the five optimization-based methods, presented in Table 3.2, returned three feasible EFM weight solutions across all eleven solvers. Of the five objective functions used, minimizing the two QP objective functions, across all solvers, returned only the Markovian solution. However, the LP-based programs returned three solutions across all eight solvers, including the Markovian weights. Given all EFMs were of the same length, neither LP objective function constrained the solution space and could have returned an infinite set of decimal-valued EFM weights. The MILP method was the only one that failed to return the Markovian solution and returned one of two feasible solutions that minimized the number of active EFMs.

The EFM probabilities were next computed for the second example network shown in Figure 3.1b. The EFM probabilities for Figure 3.2b were $P_k = [0.17, 0.17, 0.17, 0.17, 0.11, 0.11, 0.043, 0.043]$ which, when rescaled by proportionality constant α, yielded $w_k = [1.6, 1.6, 1.6, 1.6, 1.0, 1.0, 0.4, 0.4]$. This Markovian solution differed greatly from all five optimization-based methods presented in Table 3.3. Of the seven feasible solutions returned across all objective functions, none were the Markovian solution and methods that returned only a single solution were different from each other. Collectively, these results demonstrated that the Markovian constraint identified EFM weights that were unique and distinct from the other optimization-based methods, which returned multiple feasible solutions depending on the objective function and solver.

3.5.2 Application to a sphingolipid network

I next compared the Markov model and optimization-based methods to assign EFM weights in an empirically-validated network of sphingolipid metabolism of non-specific human tissue across nine subcellular compartments.[118] Figure 3.3 shows the model which consists of 69 reactions across 39 lipid classes and kinetic parameters for the healthy and Alzheimer's disease condition. The within-compartment reactions were modelled by Michaelis-Menten kinetics, while inter-compartment transport reactions were modelled by mass-action kinetics. This sphingolipid network is an ideal unimolecular model because reactions either move a lipid class from one compartment to another or transform one lipid class into another by the addition or removal of an N-acyl fatty acid or phosphocholine head group. Although Figure 3.3 suggests an open-loop network, source reactions 64-68 and sink reaction 6 contained zero-valued kinetic parameters in both conditions and therefore no flux into or out of the system. The kinetic parameter for source reaction 1 was zero in the healthy condition and set to zero (from $V_m = 0.04$) in the disease condition to ensure a closed-loop network. Altogether, there were 62

Table 3.2: EFM weights from Figure 3.1a with solvers shown below.

EFM	Method									
	Markov	Min L2	Max qSPA	Max lSPA			Min lSPA			Min milAP
1	1	1[a]	1[b]	2[c]	0[d]	1[e]	2[f]	0[g]	1[h]	2[i]
2	1	1	1	0	2	1	0	2	1	0
3	1	1	1	2	0	1	2	0	1	2
4	1	1	1	0	2	1	0	2	1	0

[a] COSMO, OSQP; [b] COSMO, OSQP; [c] CDDLib; [d] Gurobi, SCIP, GLPK; [e] OSQP, ECOS, ProxSDP, Tulip; [f] CDDLib; [g] Gurobi, SCIP, GLPK; [h] OSQP, ECOS, ProxSDP, Tulip; [i] Gurobi.

Table 3.3: EFM weights from Figure 3.1b with solvers shown below (split over two parts).

EFM	Method							
	Markov	Min L2	Max qSPA	Max lSPA				
1	1.6	0.67[a]	1.33[b]	0.67[c]	1[d]	1[e]	0[f]	1.12[g]
2	1.6	0.67	1.33	0.67	1	1	0	1.12
3	1.6	0.67	1.33	0.67	0	1	0	1.12
4	1.6	0.67	1.33	0.67	0	1	0	1.12
5	1.0	1.00	1.00	1.00	0	1	1	1.00
6	1.0	1.00	1.00	1.00	2	1	1	1.00
7	0.4	1.33	0.67	1.33	2	1	2	0.88
8	0.4	1.33	0.67	1.33	1	1	2	0.88

[a] COSMO, OSQP; [b] COSMO, OSQP; [c] OSQP, ECOS, ProxSDP; [d] Gurobi, CDDLib; [e] SCIP; [f] GLPK; [g] Tulip;

EFM	Method							
	Markov	Min lSPA						Min milAP
1	1.6	0.67[h]	1[i]	1[j]	0[k]	0[l]	1.12[m]	0[n]
2	1.6	0.67	1	1	0	0	1.12	0
3	1.6	0.67	0	1	1	0	1.12	0
4	1.6	0.67	0	1	1	0	1.12	0
5	1.0	1.00	0	1	2	1	1.00	1
6	1.0	1.00	2	1	0	1	1.00	1
7	0.4	1.33	2	1	1	2	0.88	2
8	0.4	1.33	1	1	2	2	0.88	2

[h] OSQP, ECOS, ProxSDP; [i] Gurobi; [j] SCIP; [k] CDDLib; [l] GLPK; [m] Tulip; [n] Gurobi.

remaining reactions involving 37 lipid classes.

The full ordinary differential equation (ODE) models provided in Table A.3 by Wronowska et al..[118] were reproduced with minor corrections and solved to identify the steady state lipid concentrations in both conditions (see Tables A.4 and A.5). Figure 3.3 shows the steady state fluxes computed from the kinetic parameters and steady state

lipid concentrations of both models.

I next enumerated the EFMs in the sphingolipid network using the CHMC method which returned 55 pathways. To validate the correctness of the CHMC, I also enumerated the EFMs using FluxModeCalculator[60] which is the fastest EFM enumeration program in literature. I found both methods enumerated the exact same 55 EFMs with comparable run times (1.6 ms with FluxModeCalculator and 4.3 ms with the CHMC). Of the 55 network EFMs, 11 of them were short, two-step reaction cycles such as ceramide to sphingosine and back to ceramide in the endoplasmic reticulum. Others were longer, including multiple lipid transformations and/or transport between different subcellular locations. The three longest EFMs were 14 steps long and spanned 6 subcellular compartments, although not necessarily the same ones (see Appendix A.4). I used my Markovian method as well as the optimization methods described in Table 3.1 to calculate EFM weights for the healthy and Alzheimer's disease fluxes.

Figure 3.4 shows the EFM weights across all tested methods for the healthy and disease conditions, where all methods, including mine, returned quite different solutions. The greatest agreement in EFM weights, across methods and solvers, was observed for the top 9 EFM weights in the healthy condition of Figure 3.4a. These EFMs explained 85.8% of the total network fluxes and consisted nearly exclusively of two-step reversible reactions between lipid classes within a compartment. Their weights were consistent across methods because the large fluxes between pairs of reversible reactions can only be explained by these two-step pathways. Beyond these reversible reactions, there was less agreement for the remaining EFMs consisting of longer reactions which varied by up to three orders of magnitude between solvers and across objective functions. As an increase in one EFM weight must be balanced by a decrease in another to satisfy flux conservation, EFMs assigned very large values must force others towards zero. Across all optimization-based methods and solvers, I observed a 32% probability of assigning a zero-valued weight to each of the 55 EFMs. Different objective functions displayed

varying tendencies towards sparse solutions with the MILP and maximizing shortest pathway activity by QP functions assigning the most zero-valued weights on average (29 and 26.5). Both linear programming methods that either maximized or minimized the shortest pathway activity assigned the same average number of zero-valued weights in the healthy condition.

These observations were consistent in the Alzheimer's disease condition in Figure 3.4b with high agreement for the same top 9 EFM weights shared in the healthy condition. These EFMs explained 86.7% of the total network flux with greater disagreement within solvers and across methods for the remaining EFMs. Minimizing the Euclidean norm (L2 norm), for instance, yielded weights that were up to four magnitudes smaller than the Markov solution or other objective functions. This variability corresponded to a 35% probability of any optimization-based method/solver assigning a zero-valued weight. In contrast, my Markov solution is not based on sparsity or parsimony, so all EFMs receive some non-zero weight, regardless of their length or how much flux they explain in the network. I observed that the largest 13 EFM weights explained 95% of the fluxes in both conditions and that 11 of them were shared between healthy and Alzheimer's disease (see Appendix A.5). One of the two largest EFMs unique to the healthy and Alzheimer's disease were two-step reactions cycling sphingosine and sphingosine-1-phosphate in the outer membrane and sphingomyelin to ceramide in the nucleus, respectively. The other largest EFM unique to the healthy condition consisted of five reactions that carried flux through the nucleus, cytoplasm, and endoplasmic reticulum, while the largest EFM unique to Alzheimer's disease also consisted of five reactions but through the mitochondria, cytoplasm and endoplasmic reticulum. I observed that EFM 11 (Figure 3.4b, x-axis index 30) showed the largest (3 orders of magnitude) increase in weight between Alzheimer's disease and the healthy condition. This EFM corresponded to a two-step cycle between sphingomyelin and ceramide in the nucleus which explained why its weight showed high agreement across

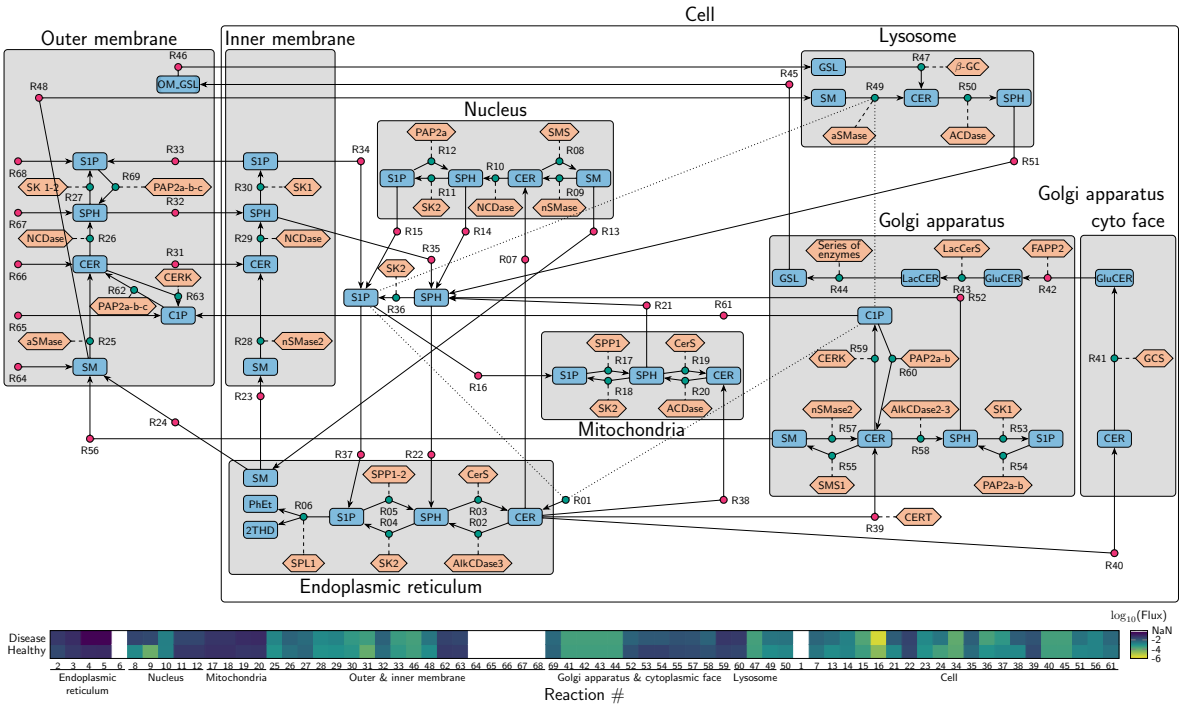


Figure 3.3: Sphingolipid kinetic model adapted from Wronowska et al.[118] Oval boxes denote lipid classes and hexagonal boxes denote enzymes. Teal circles denote transport reactions modelled by mass-action kinetics. Magenta circles denote metabolic reactions modelled by Michaelis-Menten kinetics. Solid lines connect metabolites and reactions, dashed lines denote the reaction enzyme, and dotted lines denote non-competitive inhibition.

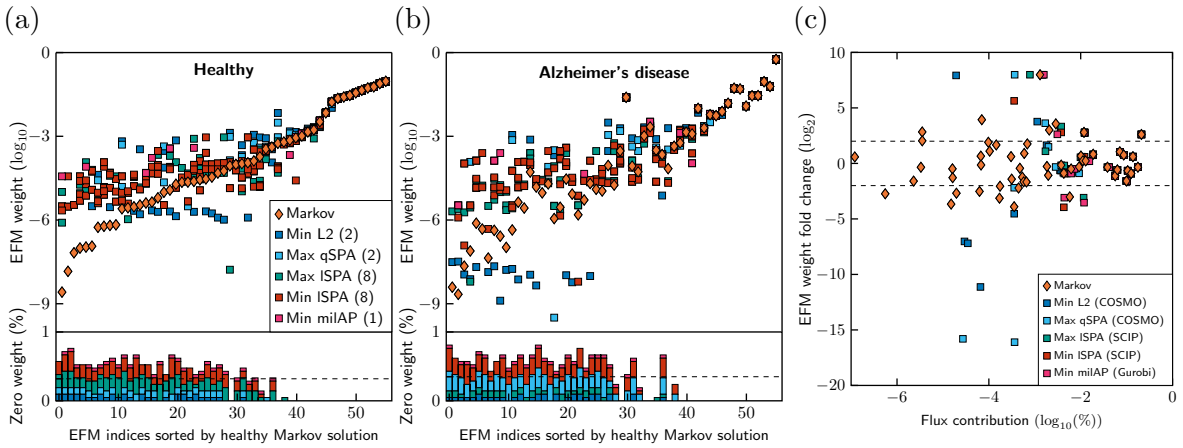


Figure 3.4: EFM weights under the Markov constraint versus five objective functions from optimization-based methods across all solvers for the (a) healthy, (b) disease condition, (c) fold change with respect to healthy with a representative solver from each method The optimization-based method codes are described in Table 3.1. Numbers in parentheses correspond to the solvers used for each mathematical optimization. Histogram bars show the frequency of zero-valued EFM weights from the optimization-based methods normalized across all 21 objective functions and solvers.

both Markov and optimization methods.

Given the majority of network fluxes across both conditions were explained by the same 11 EFMs, I next investigated whether metabolic differences were explainable through other differentially active EFMs. Figure 3.4c shows the fold change between the Alzheimer's disease and healthy condition EFM weights as a function of the total explained flux in the healthy condition. Weights were selected from the objective function solver that best reconstructed the individual and total fluxes across both networks (see Appendices A.6 and A.7). The top 9 EFM weights explaining the majority of the network fluxes are located on the right-most cluster of Figure 3.4c where there is high agreement across all methods. Only one of those EFMs showed over four-fold increase in the Alzheimer's disease versus healthy condition. This two-step EFM pathway corresponded to the remodelling of sphingosine and sphingosine-1-phosphate in the endoplasmic reticulum. These data suggest the sphingolipidome in Alzheimer's disease is altered through pathways carrying smaller units of steady state flux. According to the Markov solution, I observed extreme fold changes ($|\log_2 FC| \geq 4$) for 17 EFMs that explained between 10^{-6} and 0.1% of the total network flux. In contrast, there were fewer fold changes for the optimization-based methods due to the sparsity of EFM weights for the disease and healthy conditions. Between 25 and 34 undefined fold changes were observed across all five optimization methods because these methods assigned a zero-valued EFM weight to one or both conditions. Furthermore, the optimization methods also showed extreme fold changes for EFMs that explained far less flux than my Markov solution. Of the 6 to 10 extreme fold changes observed across all LP/MILP/QP methods, their EFM weights only explained between 0.0001% to 0.001% of the total healthy fluxes. Both linear and mixed-integer programming methods showed between 6-10 extreme fold changes among EFMs that explained $\approx 0.001\%$ of the total healthy flux. Although the quadratic programming methods showed 8 and 10 extreme fold changes, those EFMs explained $\approx 0.0001\%$ of the total healthy flux.

3.6 Conclusion

In this chapter, I explored the computational problem of identifying EFM weights explaining steady state fluxes in a metabolic network. Although this problem has mainly been addressed by optimization-based methods, I showed how different objective functions may nonetheless fail to properly constrain the solution space of EFM weights. I then proposed a Markovian constraint, in which EFM weights are made proportional to the probability of a metabolite "particle" travelling along different paths through the network. My Markovian approach results from a natural, biophysical assumption, which may offer greater accuracy to real metabolic networks.

I applied my Markov method to identify differentially active EFM weights in a model of sphingolipid metabolism in Alzheimer's disease and compared the results against optimization-based methods. While all methods strongly agreed on the top 9 (16%) EFM weights, the Markov and objective functions across all solvers greatly differed for the remaining EFMs. Upon inspection, I found that EFMs with the largest weights and consensus across flux decomposition methods corresponded to two-step remodelling cycles between pairs of lipid classes, while EFMs with smaller weights typically involved more reactions and/or spanned compartments.

While I developed my CHMC method to uniquely estimate EFM weights from metabolic flux networks, I highlight algorithmic applications for other network flow problems in biology. Ion channel behaviour or protein folding dynamics may be characterized by the cyclic flow of conformational changes of the atom positions and momenta,[181,211] and these simple cycle probabilities may be uniquely computed by my CHMC method. Flow decomposition ambiguity also occurs in ribonucleic acid (RNA) quantification algorithms where the same set of reads mapped to a splice graph may be explained by more than one set of transcript isoforms, even under parsimonious constraints.[212,213]

There exist several avenues of future work to extend my Markov chain model to identify EFM weights. First, my method is only applicable to metabolic networks of unimolecular reactions and my methodology must be expanded to account for networks containing higher order reactions. Second, algorithmic developments are required to scale the Markov method to genome-wide networks containing millions of EFMs. While the Julia code I provide can compute two to three thousand EFM weights in seconds, enumerating the CHMC state space is prohibitively memory intensive in larger-scale networks. Third, I anticipate my work will encourage differential analysis of EFM weights to quantify metabolic changes across conditions. Previous analyses have typically computed EFM weights using a variety of objective functions and used a consensus-based approach to identify active EFMs.[53,55] As my method uniquely identifies EFM weights, I foresee a deeper analysis of EFM weights and their incorporation in statistical learning models to study differential metabolism.

4

Computing elementary flux mode weights by cycle-history Markov chains in discrete- and continuous-time

4.1 Author contributions

Justin G. Chitpin developed the methods and wrote this chapter with guidance from Theodore J. Perkins.

4.2 Abstract

The cycle-history Markov chain (CHMC) is a newly proposed method for uniquely computing the elementary flux mode weights that reconstruct the steady state fluxes of

a unimolecular reaction network. In its original formulation, this method is a discrete-time Markov chain (DTMC) with transition probabilities proportional to the outgoing fluxes. One may wonder what would happen if one imagined that particle jumping between between metabolites in continuous rather than discrete time. This is more biophysically realistic than assuming that particle transitions occur over discrete time steps. It also raises the question of whether elementary flux mode (EFM) weights are different if computed from a DTMC-versus-continuous-time Markov chain (CTMC). In this work, I show that the CHMC may be defined as a CTMC following a mass-action kinetic model of flux. I further show both discrete and continuous-time CHMC compute identical EFM weights by proving their arc probabilities are proportional to each other.

4.3 Introduction

The elementary flux mode (EFM) is defined as a minimal pathway that can support steady state flux through a metabolic network.[37] Given a metabolic network and its steady state fluxes, a fundamental problem in EFM literature is to decompose the observed network fluxes into EFM weights.[210] In Chapter 3, I introduced the cycle-history Markov chain (CHMC) which is the first method that can uniquely assign EFM weights for any unimolecular metabolic network under steady state.[182] This CHMC models a single particle moving through a metabolic network as a discrete-time Markov chain (DTMC) with nodes representing distinct metabolites and edges proportional to the outgoing steady state fluxes. The CHMC is constructed in such a way as to trace the long-term probabilities of that particle transiting a sequence of metabolite states, or equivalently a simple cycle, corresponding to each EFM.

In this chapter, I approach the construction of the CHMC from a continuous-time Markov chain (CTMC). Starting with a unimolecular flux network, I first model these reactions as differential equations following first order mass-action kinetics. I then show

how these equations describing any unimolecular flux network can be reformulated in terms of an equivalent CTMC with particle jump times based on the mass-action kinetic parameters. I then show that the steady state arc probabilities in this CTMC and corresponding embedded DTMC are proportional to each other. This leads to a final proof where I show that the EFM weights computed from the continuous-time or discrete-time CHMC are identical. Thus, one could define the continuous-time version of the CHMC in Chapter 3, yet the EFM weights would remain the same.

4.4 Metabolic flux as a single particle, CTMC

Consider the steady state flux in a closed, connected, unimolecular metabolic network. Let A_i be the abundance of metabolite i and let R_{ij} be the reaction rate (a kinetic parameter in this context not to be confused with a flux) transforming metabolite i into metabolite j. Let $A = \sum_i A_i$ be the total abundance of all molecules. Lastly, let the relative abundance of molecule i be $N_i = A_i/A$. The ordinary differential equation (ODE) describing the abundance of molecule i is

$$\frac{dA_i}{dt} = \sum_{j \neq i} R_{ji} A_j - \sum_{j \neq i} R_{ij} A_i, \tag{4.1}$$

and the corresponding equation for the normalized abundance of molecule i is

$$\frac{d}{dt}\left(\frac{A_i}{A}\right) = \frac{dN_i}{dt} = \sum_{j \neq i} R_{ji} N_j - \sum_{j \neq i} R_{ij} N_i. \tag{4.2}$$

The normalized differential equations define a continuous-time stochastic process known as a CTMC.[179] This CTMC models the process of single substrates undergoing reactions to remodel into single products. In Markov chain parlance, these metabolites are referred to as states while the reactions between substrate-product pairs are referred to as transitions. A fundamental property of this Markov chain is that the process of one

substrate transforming into a product is independent of previous biochemical reactions leading to that reaction. This is known as the Markov constraint where the probability of transitioning from one state to another only depends on the present state.

In a CTMC, the time between state transitions is modelled as an exponentially distributed random variable. The rates of each transition define a generator matrix Q with rate constants $q_{ij} = R_{ij}$ for $i \neq j$ and $q_{ii} = -\sum_{j \neq i} R_{ij}$. If every state is reachable from another state, there exists a unique, steady state probability P_i of observing state i in the chain. The time-dependent evolution for state i is described by Kolmogorov's forward equations:

$$\frac{dP_i}{dt} = \sum_{j \neq i} q_{ji} P_j + q_{ii} P_i. \tag{4.3}$$

Note the similarity between Kolmogorov's forward equation and the normalized system of ODEs describing the unimolecular flux network. The steady state probability P_i is equivalent to the normalized, steady state concentration of metabolite i. The flux along each Markov chain arc is therefore proportional to the mass-action kinetic fluxes of the corresponding metabolic network.

To confirm that this Markov chain captures the steady state flux constraint, consider the flux from molecule i to j in the metabolic network $F_{ij} = R_{ij} \times A_i$. The normalized flux is $G_{ij} = R_{ij} \times \frac{A_i}{A} = R_{ij} \times N_i$. If the Markov chain is at steady state

$$\frac{dP_i}{dt} = 0 = \sum_{j \neq i} q_{ji} \times P_j + q_{ii} \times P_i = \sum_{j \neq i} R_{ji} \times N_j - \sum_{j \neq i} R_{ij} \times N_i \tag{4.4}$$

where the normalized molecular concentration is equal to the steady state probability of state i in the chain ($N_i = P_i$). This leads to

$$\sum_{j \neq i} G_{ij} = \sum_{j \neq i} G_{ji}, \tag{4.5}$$

which, when both sides are multiplied by A, shows that the steady state forward equa-

tions of the Markov chain satisfy flux conservation.

Next, let us define the individual steady state arc fluxes for this CTMC. These arc fluxes are defined as $F^c_{ij} = q_{ij} \cdot P_i$ which represents the long-term proportion of time a particle spends in state i before jumping to state j. A similar set of steady state arcs can be defined for the corresponding embedded DTMC. Denoting T as the corresponding transition probability matrix and p as the steady state probabilities, the arc probability fluxes are defined as $F^d_{ij} = T_{ij} \cdot p_i$. These arc fluxes reflect the long-term probability of a particle occupying state i before transitioning to state $j \neq i$.

4.5 Steady state arc probabilities are proportional between the CHMC and embedded DTMC

Theorem 1. *In a CTMC, the steady state arc fluxes are proportional to the steady state arc probabilities of the corresponding embedded DTMC.*

Proof. Let Q be the generator matrix of a CTMC, with Q_{ij} being the transition rate from state i to state j. In matrix form, this can be written as:

$$Q = \begin{bmatrix} Q_{11} & Q_{12} & \cdots & Q_{1n} \\ Q_{21} & Q_{22} & \cdots & Q_{2n} \\ \vdots & \vdots & & \vdots \\ Q_{n1} & Q_{n2} & \cdots & Q_{nn} \end{bmatrix} \tag{4.6}$$

where n is the number of states in the chain. Note that $Q_{ii} = -\sum_{j \neq i} Q_{ji}$. Under the ergodicity assumptions, a CTMC has a unique steady state distribution $q = [q_1, q_2, \ldots, q_n]$. This steady state satisfies $q \cdot Q = \mathbf{0}$, where $\mathbf{0}$ denotes a length-n vector of zeros.

The embedded DTMC has transition matrix T, with T_{ij} being the transition probability from state i to state j. The matrix entries are defined as:

$$T_{ij} = \begin{cases} -Q_{ij}/Q_{ii} & \text{for } i \neq j \\ 0 & \text{for } i = j \end{cases} \tag{4.7}$$

Thus, T is:

$$T = \begin{bmatrix} T_{11} & T_{12} & \cdots & T_{1n} \\ T_{21} & T_{22} & \cdots & T_{2n} \\ \vdots & \vdots & & \vdots \\ T_{n1} & T_{n2} & \cdots & T_{nn} \end{bmatrix} = \begin{bmatrix} 0 & -Q_{12}/Q_{11} & \cdots & -Q_{1n}/Q_{11} \\ -Q_{21}/Q_{22} & 0 & \cdots & -Q_{2n}/Q_{22} \\ \vdots & \vdots & & \vdots \\ -Q_{n1}/Q_{nn} & -Q_{n2}/Q_{nn} & \cdots & 0 \end{bmatrix} \tag{4.8}$$

With the ergodicity assumptions, the DTMC has a unique, steady state distribution $p = [p_1, p_2, \ldots, p_n]$ which satisfies $p \cdot T = p$. Equivalently, $p \cdot (T - I) = 0$. Defining $P = T - I$, it is equivalent that $p \cdot P = 0$. Note that:

$$P = \begin{bmatrix} -1 & -Q_{12}/Q_{11} & \cdots & -Q_{1n}/Q_{11} \\ -Q_{21}/Q_{22} & -1 & \cdots & -Q_{2n}/Q_{22} \\ \vdots & \vdots & & \vdots \\ -Q_{n1}/Q_{nn} & -Q_{n2}/Q_{nn} & \cdots & -1 \end{bmatrix} \tag{4.9}$$

Let D be the diagonal matrix with values $-Q_{11}, -Q_{22}, \ldots, -Q_{nn}$ down the main diagonal.

$$D = \begin{bmatrix} -Q_{11} & 0 & \cdots & 0 \\ 0 & -Q_{22} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & -Q_{nn} \end{bmatrix} \tag{4.10}$$

Observe that $D \cdot P = Q$. Since $q \cdot Q = 0$, it must follow that $q \cdot D \cdot P = 0$. This

means the vector $q' = q \cdot D$ is the left nullspace of matrix P. That nullspace also contains the steady state distribution p because $p \cdot P = 0$ which is one dimensional because of the ergodicity assumptions. Otherwise, there could be multiple distinct, linearly independent steady state distributions. Therefore, the vector p and the vector $q' = q \cdot D = [-Q_{11}q_1, -Q_{22}q_2, \ldots, -Q_{nn}q_n]$ must be proportional. That is, there exists a real constant z such that $p = z \cdot q'$.

With these preliminaries, Theorem 1 can now be proven by examining the steady state arc fluxes of the CTMC and embedded DTMC. Let $F_{ij}^c = q_i \cdot Q_{ij}$ denote the steady state arc fluxes between state i and state $j \neq i$ for the CTMC and let $F_{ij}^d = p_i \cdot P_{ij}$ denote the steady state arc fluxes between state i and state $j \neq i$ for the embedded DTMC. Noting that $p = z \cdot q'$, so $p_i = z \cdot q_i'$, $P_{ij} = -Q_{ij}/Q_{ii}$, and $q_i' = q_i \cdot (-Q_{ii})$:

$$\begin{aligned} F_{ij}^d &= p_i \cdot P_{ij} \\ &= z \cdot q' \cdot P_{ij} \\ &= z \cdot q' \cdot \frac{-Q_{ij}}{Q_{ii}} \\ &= z \cdot q \cdot (-Q_{ii}) \cdot \frac{-Q_{ij}}{Q_{ii}} \\ &= z \cdot q \cdot Q_{ij} \\ F_{ij}^d &= z \cdot F_{ij}^c \end{aligned}$$

Corollary 1.1. Because the previous theorem applies to any CTMC and its embedded DTMC, it applies, in particular, to the CHMC. Thus, the EFM probabilities computed from the CHMC will be identical whether the transition matrix consists of probabilities proportional to steady state network fluxes or a set of matching kinetic parameters.

4.6 Conclusion

In this chapter, I showed how the EFM probabilities computed from the CHMC are identical under either a CTMC or its embedded DTMC. In a discrete-time CHMC, the probability of each EFM can be thought of as the long-term proportion of occurrences of a single particle transitioning through the corresponding state sequences. This interpretation changes somewhat is a continuous-time CHMC, where a particle dwells on a given metabolite following an exponential distribution with rate equal to the sum of its kinetic parameters. That particle then jumps to a metabolic product with probability proportional to that kinetic parameter. The EFM probabilities in continuous-time can be interpreted as the proportion of time spent transiting a given EFM. Intuitively, increasing the kinetic parameter associated with a reaction along a given EFM should increase the probability of the particle transiting that EFM. However, particle dwell times decrease exponentially as the kinetic parameters increase and shorten the proportion of time spent transiting that EFM. These opposing time differences "balance out" in the context of EFM probabilities. While it may be more intuitive to compute the CHMC in continuous-time due to the clear relationship with mass-action kinetic equations, the discrete-time version is generally more convenient because it requires steady state fluxes as input rather than a full mass-action kinetic model describing the metabolic network.

5

A dominant pathway flux hypothesis

Justin G. Chitpin developed the methods, conducted the experiments, analyzed the results, and wrote this chapter with guidance from Theodore J. Perkins.

5.1 Abstract

Metabolic networks are a highly complex organization of biochemical reactions required to sustain life. Understanding how reaction rates are distributed in these networks is fundamental for understanding how organisms adapt metabolically to external perturbations and to optimize microbial factories to produce desirable biochemical compounds. Existing literature has shown that both the structure of these networks and their metabolic fluxes follow power law distributions. This uneven distribution of flux, in particular, raises important questions regarding the distribution of pathway fluxes

through these networks. Here, I ask whether pathway fluxes, operationalized as elementary flux modes (EFMs), exhibit this unevenness in a sphingolipid kinetic model of metabolism. I find that both fluxes and their corresponding EFM weights exhibit dominance characterized by a log-normal, rather than power law, distribution and empirically demonstrate that a small subset of EFMs explain the majority of flux in this network (see chapter 8).

5.2 Introduction

Cellular metabolism consists of a highly complex network of enzyme-catalysed and transport reactions that interconvert and shuttle thousands of metabolites within and across subcellular compartments.[1] Computational pipelines leveraging large-scale sequencing data are routinely used to reconstruct thousands of genome-scale metabolic models (GEMs) across diverse organisms.[69,71,214] These networks are commonly studied by flux balance analysis (FBA) and kinetic models to understand how flux is routed through metabolic pathways under biological conditions of interest.[41,85,215]

While much effort has been made to characterize the topology of metabolic networks, many questions remain regarding the organization of the steady state fluxes within these systems. Previous studies have shown that many metabolic networks are scale-free, implying that the degree distribution of metabolites follows a power law.[216,217] Almaas et al. further showed that random FBA solutions for a GEM of *E. coli* K12 MG1655 also exhibited power law distributions, with the majority of reactions carrying small fluxes contrasting with a minority of those carrying large fluxes.[218] Their results thus suggest that the distribution of steady state fluxes is independent of both network size and metabolic phenotype.

Understanding how steady state reaction fluxes are distributed in metabolic networks raises a similar question regarding the organization of pathway fluxes. In the

present context, steady state pathway flux refers to the stoichiometrically-balanced flow of metabolites through a network (i.e. a source-to-sink pathway) or within a network (i.e. an internal cycle). Intuitively, an uneven distribution of pathway fluxes can be explained by an uneven distribution of reaction fluxes. This would imply that individual reactions carrying high flux are linked together through common pathway fluxes. This idea leads me to to formulate a *dominant pathway flux hypothesis*, whereby the majority of steady state reaction fluxes in a network are explained by a small subset of pathway fluxes.

Investigating this dominant pathway flux hypothesis requires a metabolic pathway definition that can account for the observed flux in any metabolic network under steady state. The EFM is therefore useful since they are defined as steady state pathways whose positive linear combinations can reconstruct any set of steady state fluxes in a metabolic network.[37,210] Furthermore, unique EFM flux decompositions are now possible for the special case of unimolecular flux networks through the cycle-history Markov chain (CHMC) method proposed in Chapter 3.[182]

In this chapter, I leverage my CHMC method to characterize the distribution of reaction fluxes and EFM pathway fluxes in a modified sphingolipid kinetic model of Alzheimer's disease and control patients.[118] Specifically, I introduce gene expression values taken from publicly available datasets which are used to estimate sample-specific reaction fluxes and EFM weights for individual patients. From these values, one could either generate a distribution of fluxes or weights across individuals (looking at how the same flux or EFM weight changes) or within individuals (looking at how all fluxes or EFM weights change). I investigate the latter and ask what is the form of these flux and EFM weight distributions. I ask whether these distributions approximately follow a power law, log-normal, or exponential distribution. The first two distributions are chosen since they often arise when describing natural phenomena[219] and biological data. These distributions are considered heavy-tailed since their tail approaches zero

much faster than the light-tailed exponential distribution.[220]

I first analyse the unperturbed kinetic model to show that both reaction fluxes and EFMs weights within each condition approximately follow a log-normal distribution. I then show that this log-normal distribution of reaction fluxes and EFM weights is maintained in the sample-specific sphingolipid kinetic models within individuals of both conditions. Using additional synthetic flux data generated by log-uniformly sampling enzyme expression coefficients, I further show that log-normal pathway flux dominance is largely independent of the distribution of $V_{\max}$ kinetic parameters with particular EFMs tending to dominate over others. These results motivate a final analysis of a class of theoretical networks to explain the observed log-normal flux dominance in the sphingolipid kinetic model and potentially in CHMC analyses of other GEMs.

5.3 Methods

5.3.1 Mapping gene expression data to sphingolipid model reactions

Bulk ribonucleic acid (RNA)-sequencing samples were taken from Srinivasan et al.[134] from `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE125583`. These samples corresponded to outer brain sections of the fusiform gyrus across 219 Alzheimer's disease patients and 70 age-matched controls.

The genes reported by Wronowska et al. in their sphingolipid kinetic model[118] were manually translated into NCBI Entrez IDs. These genes correspond to enzyme-catalysed reactions in the sphingolipid model. For some reactions, a single gene encodes the corresponding enzyme while for other reactions multiple genes may encode isoforms of the same enzyme (see Appendix B.2). According to gene-protein-reaction (GPR) rules, the expression levels of these isoforms are summed together to compute an over-

all enzyme expression level. The enzyme expression coefficient of Michaelis-Menten reaction j in sample i is therefore defined as

$$E_{i,j} = g_{i,j} \cdot \left(\prod_{i=1}^{n} g_{i,j} \right)^{\frac{-1}{n}}, \tag{5.1}$$

where g_j is the summed transcripts per million (TPM) of genes mapping to reaction j and n is the number of samples within each condition. The geometric mean was specifically chosen to account for large variations in gene counts across individuals.

5.3.2 Modifying to the sphingolipid kinetic model

The sphingolipid kinetic model used in this study was constructed by Wronowska et al.[118] and reproduced with minor corrections.[182] The model consists of 62 reactions containing non-zero flux across 37 lipid classes. Since the flux directions are identical across the Alzheimer's disease and control conditions, the CHMC enumerated 55 total EFMs. But with sample-specific enzyme expression perturbations, any change in flux direction along reversible transport reactions would alter the network structure and therefore change the number of EFMs. To address this, all reversible mass-action transport reactions were modelled as pairs of irreversible reactions. This procedure increases the number of EFMs to 267 but guarantees the same number of EFMs, regardless of the sample-specific flux distribution within each condition.

5.3.3 Solving the kinetic models and computing EFM weights

All calculations were performed in Julia version 1.10.2. Steady state lipid class concentrations were numerically solved using the Rosenbrock-Wanner-Wolfbrandt method[221] (implemented as Rosenbrock23 in `DifferentialEquations.jl`). The corresponding

steady state fluxes were analytically solved and EFM weights computed using my program `MarkovWeightedEFMs.jl` version 2.0.2, available at `https://github.com/jchitpin/MarkovWeightedEFMs.jl`.

5.3.4 Distribution fitting

Distribution fitting of steady state fluxes and EFM weights was performed in the following two ways (described as method one and method two) using R version 4.4.1.

Method one

The R packages `MASS` and `fitdistrplus` were used to fit non-truncated, log-normal and exponential distributions to the steady state fluxes and rescaled EFM weights by maximum likelihood estimation (MLE). Recall that probability density function (PDF) of the exponential distribution is

$$\Pr(X = x) = e^{-\lambda x} \quad \text{for } x \geq 0 \tag{5.2}$$

with rate parameter λ, while the PDF of the log-normal distribution is

$$\Pr(X = x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(\frac{-(\ln x - \mu)^2}{2\sigma^2}\right) \quad \text{for } x \geq 0 \tag{5.3}$$

with parameters μ (mean) and σ (standard deviation). The PDF of the power law takes the form

$$\Pr(X = x) = \frac{\alpha - 1}{x_{\min}} \left(\frac{x}{x_{\min}}\right)^{-\alpha} \tag{5.4}$$

with scaling parameter α and minimum threshold $x \geq x_{\min}$. The `fitdistrplus` solves these parameters numerically using the Nelder-Mead and Broyden–Fletcher–Goldfarb–Shanno (BFGS) method implemented in the `stats` package, although it should be noted that

the parameters could be fitted analytically by solving the MLE. The power law distributions were fit manually since this distribution is not provided in the R packages above. The scaling parameter k was computed by method of MLE, yielding

$$\hat{k} = 1 + N\Big(\sum_{i=1}^{N}\Big[\ln\Big(\frac{x_i}{x_{\min}}\Big)\Big]\Big)^{-1}, \tag{5.5}$$

with $x_{\min}$ set to the smallest value in each data sample (see Appendix B.1 for the derivations).

Model fits were assessed visually and quantitatively using the `fitdistrplus` package. The package includes functions to graph the quantile-quantile (Q-Q), probability-probability (P-P), and cumulative distribution function (CDF) plots to qualitatively inspect the distribution fits to the flux or EFM weights. The Q-Q plots graph the empirically observed fluxes or EFM weights on the y-axis with the fitted model evaluated at those values and displayed on the x-axis. Similarly, the P-P plots graph the empirical CDF of the observed fluxes or EFM weights on the y-axis with the fitted model CDF evaluated at those values and displayed on the x-axis. For both Q-Q and P-P plots, a strong model fit is depicted by coordinates that fall around the $y = x$ diagonal.

To give an example of the Q-Q, consider the set of values $x = [1, 3, 4, 5, 7]$, which I may view as roughly normally distributed with sample mean $\mu = 4$ and sample variance $\sigma^2 = 5$. Evaluating this normal distribution at the x values gives fitted values $\hat{x} = [0.072, 0.16, 0.18, 0.16, 0.072]$. The coordinates $(x_i, \hat{x}_i)$ are thus plotted on the Q-Q plot. Regarding the P-P plot, the empirical cumulative probabilities for x are $p = [0.2, 0.4, 0.6, 0.8, 1.0]$. Evaluating the normal distribution at these x values yields fitted cumulative probabilities $\hat{p} = [0.090, 0.33, 0.5, 0.67, 0.91]$. The coordinates $(p_i, \hat{p}_i)$ are thus plotted on the P-P plot.

Finally, the CDF plots graph the empirical or model CDF values on the y-axis as a function of the fluxes or EFM weights displayed on the x-axis. Goodness-of-fit

was assessed quantitatively by log-likelihood of the fitted model and the Kolmogorov-Smirnov (KS) statistic which follows the formula $\sup|F_n(x) - F(x)|$ and was computed by

$$D = \max(D^+, D^-) \tag{5.6}$$

where $D^+ = \max_{i=1,\ldots,n}(\frac{i}{n} - F_i)$ and $D^- = \max_{i=1,\ldots,n}(F_i - \frac{i-1}{n})$.

Method two

The R package `poweRlaw`[222] was used to fit lower truncated power law, log-normal, and exponential distributions to the steady state fluxes and rescaled EFM weights by MLE. The method estimates the $x_{\min}$ for all three distributions using the approach described by Clauset et al..[223] Under this method, P-values assessing the goodness-of-fit were computed by generating 1000 bootstrap samples under the empirical distribution, fitting models to each bootstrapped sample, and computing the fraction of fits whose KS statistic is larger than the value for the original, empirical data. In this goodness-of-fit test, Clauset et al. reject the hypothesis of the data following a given distribution if $P \leq 0.1$. This same P-value threshold was used in all analyses presented in this chapter.

5.4 Results

5.4.1 A transcript-guided kinetic model of sphingolipid metabolism

Figure 5.1a visualizes the sphingolipid kinetic model which includes nine subcellular compartments with intra- and inter-compartmental reactions modelled by Michaelis-Menten and mass-action kinetics, respectively. This network was chosen because all

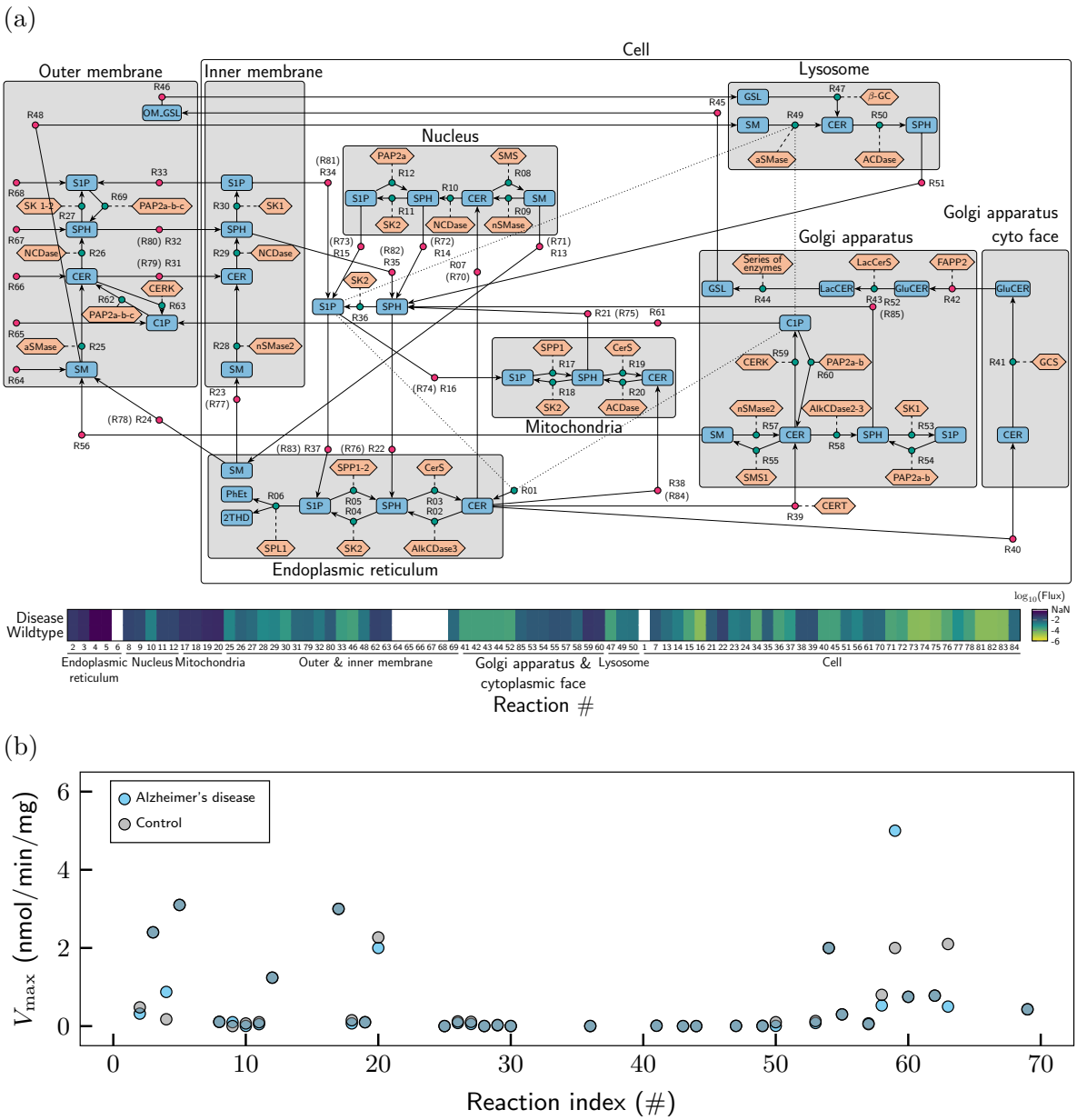


Figure 5.1: Sphingolipid kinetic model adapted from Wronowska et al.[118] (a) Oval boxes denote lipid classes and hexagonal boxes denote enzymes. Teal circles denote transport reactions modelled by mass-action kinetics. Magenta circles denote metabolic reactions modelled by Michaelis-Menten kinetics. Solid lines connect metabolites and reactions, dashed lines denote the reaction enzyme, and dotted lines denote non-competitive inhibition. Mass-action kinetic reactions are split into pairs of irreversible reactions with the reverse reaction index shown in parentheses. (b) Alzheimer's disease and control $V_{\max}$ parameters for all 35 Michaelis-Menten reactions with non-zero fluxes.

reactions carrying non-zero flux are unimolecular in nature and hence can be decomposed onto EFMs using my CHMC method described in Chapter 3. A kinetic model is

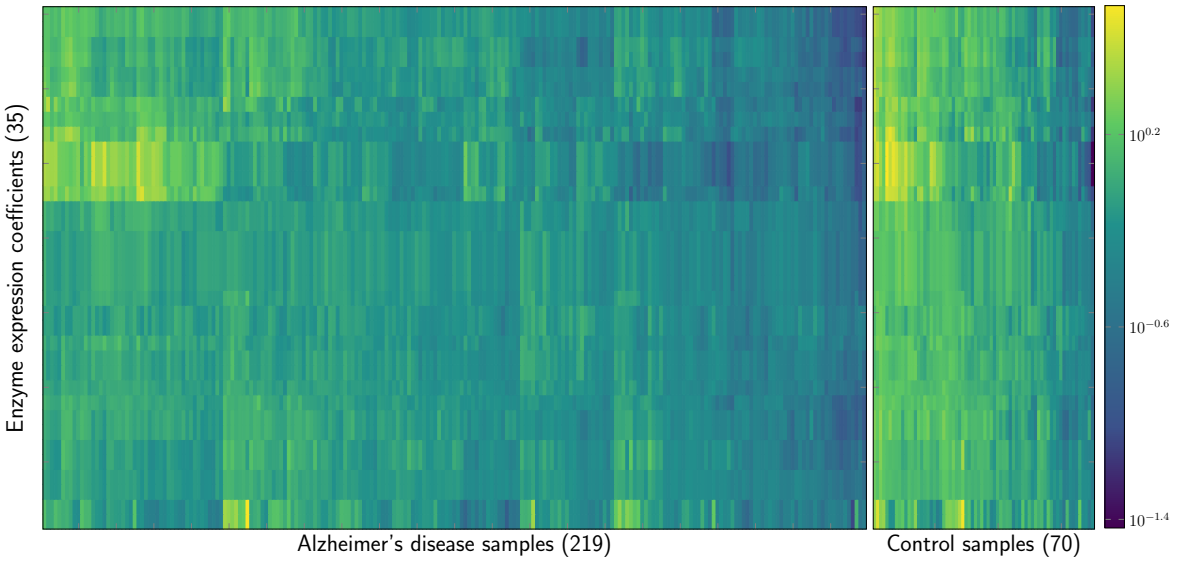


Figure 5.2: Enzyme expression coefficients for the 35 Michaelis-Menten reactions with non-zero fluxes.

also ideal since it avoids the possibility of assigning thermodynamically infeasible flux through substrate cycles which can occur when using FBA methods. The Alzheimer's disease and control condition models primarily differed in their $V_{\max}$ parameters of the Michaelis-Menten reactions (Figure 5.1b) which will become important in subsequent sections.

To obtain a sample-specific distribution of sphingolipid fluxes for the Alzheimer's disease and control conditions, the kinetic model was perturbed with bulk gene expression data acquired from fusiform gyrus sections of 219 Alzheimer's disease and 70 age-matched control patients.[134] Following GPR rules,[68] an enzyme expression coefficient was computed by normalising the TPM of genes mapping to each reaction by the geometric mean across all condition-specific samples. This resulted in 289 different sets of perturbations to the kinetic model. These enzyme expression coefficients and their variability within and across conditions are visualized in Figure 5.2. Note that enzyme expression coefficients are identical across certain rows due to identical enzymes governing these reactions.

Based on these enzyme expression coefficients, the sample-specific flux for Michaelis-

Menten reaction j was computed following

$$v_j = E_j \cdot \frac{v_{\max,j} \cdot [S]}{K_{m,j} + [S]}, \tag{5.7}$$

where E_j is the enzyme expression coefficient. The ordinary differential equation (ODE) models for each Alzheimer's disease or control sample were then solved to compute the steady state fluxes along each reaction and decomposed onto 267 EFMs under a Markovian constraint.[182]

5.4.2 Analysis of the unperturbed kinetic models

Steady state fluxes approximately follow a log-normal distribution within conditions

I first investigated whether the steady state reaction fluxes were unevenly distributed in the unperturbed sphingolipid model. Power law, log-normal, and exponential distributions were fit to the Alzheimer's disease and control model using two approaches described in the Methods section as Method one and Method two.

Statistical analysis of the steady state fluxes showed the log-normal distribution to exhibit the best fit for both the unperturbed Alzheimer's disease and control condition model, regardless of the fitting method and goodness-of-fit statistic. For simplicity, Figure 5.3 shows the non-truncated fits (of Method one) represented as Q-Q, P-P, and complementary CDF plots. The Q-Q plot visualizes the agreement between the observed steady state fluxes against the theoretical values computed from a fitted distribution. From Figure 5.3a/d, it is clear that the reaction fluxes within each condition are best-described by the log-normal distribution with the power law distribution overestimating the steady state fluxes by 1-2 orders of magnitude. Similar conclusions are drawn from the P-P plot of Figure 5.3b/e which visualizes the agreement between the

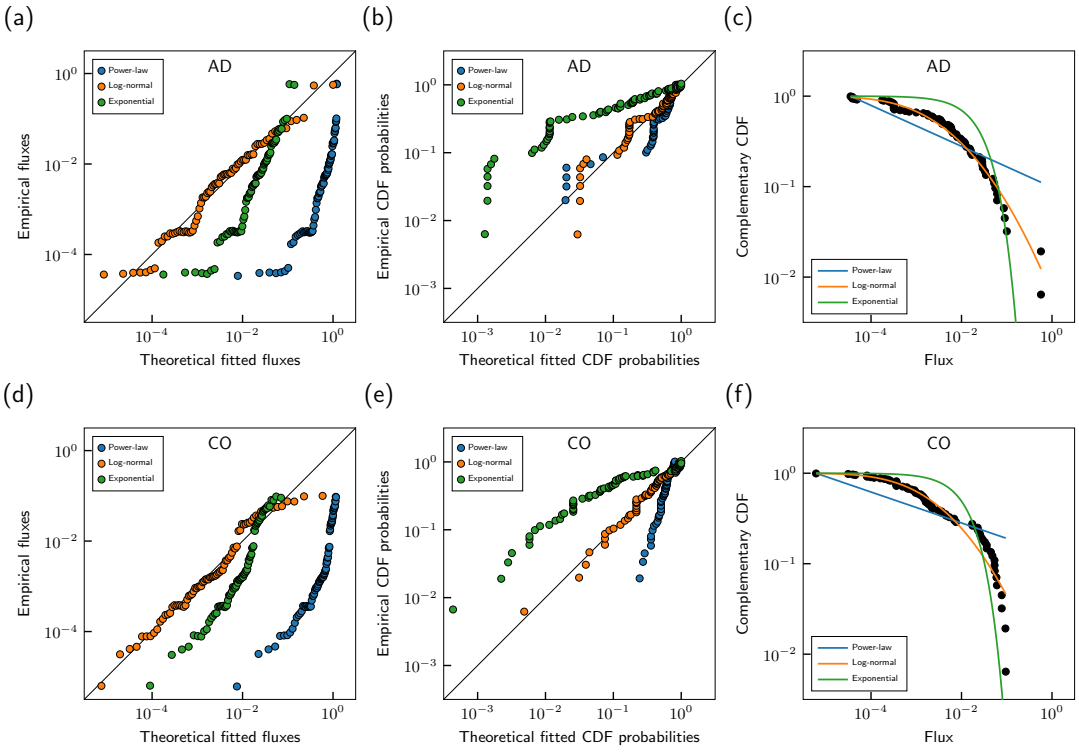


Figure 5.3: Fitted distributions to the reaction fluxes within unperturbed conditions using Method one. (a-c) Q-Q plot, P-P plot, and complementary CDF plot of Alzheimer's disease samples. (d-f) Q-Q plot, P-P plot, and complementary CDF plot of control samples.

empirical cumulative probabilities of observing a particular reaction flux against the theoretical cumulative probabilities for a given distribution. Here, the exponential distribution shows the worst fit since the cumulative flux probabilities $< 10^{-2}$ are about an order of magnitude smaller than the empirical CDF. These qualitative observations supporting the log-normal distribution are also backed up with the greatest log-likelihood of the model fit against the two other distributions in either condition. The KS statistic for the log-normal fit was also the smallest ($D_{ln}^{AD} = 0.12$) versus the power law distribution ($D_{pl}^{AD} = 0.24$) and exponential distribution ($D_{exp}^{AD} = 0.41$) in the disease condition and likewise in the control condition ($D_{ln}^{CO} = 0.11$, $D_{pl}^{CO} = 0.32$, $D_{exp}^{CO} = 0.45$). Overall, the goodness-of-fit is best shown in Figure 5.3c/f which plots the complementary CDF as a function of the empirically observed reaction fluxes. The log-normal fit accurately predicted the cumulative reaction fluxes over 4 orders of magnitude in contrast with

the power law and exponential fits. The same findings were obtained using Method two (of Clauset et al.) with only the log-normal distribution passing the goodness-of-fit test while retaining the smallest $x_{\min}$ values in the Alzheimer's disease condition. Although both exponential and log-normal distribution P-values passed the $P > 0.1$ threshold for the control condition, the log-normal fit $x_{\min}$ was 2.5 times smaller than the exponential fit and therefore the best distribution modelling the control reaction fluxes.

EFM weights approximately follow a log-normal distribution within conditions

To test for dominance of pathway fluxes, the EFM weights for the Alzheimer's disease and control models were rescaled by dividing them by the largest EFM weight within their respective condition. This normalisation leads to unitless EFM weights that have a maximum rescaled weight of one with all other weights represented as a fraction of that largest pathway. Following the previous section, power law, log-normal, and exponential distributions were fit to the rescaled EFM weights within each condition. The uneven distribution of pathway fluxes is shown in Figure 5.4 where the majority of EFM weights are orders of magnitude smaller than the most active EFM within that sample. Furthermore, the log-normal distribution exhibited the best fit among all three distributions based on the Q-Q, P-P, and complementary CDF plots in both the Alzheimer's disease and control condition. These findings were supported with the smallest KS statistics for the log-normal distributions in the Alzheimer's disease ($D_{ln}^{AD} = 0.045$, $D_{pl}^{AD} = 0.32$, $D_{exp}^{AD} = 0.84$) and control condition ($D_{ln}^{CO} = 0.043$, $D_{pl}^{CO} = 0.30$, $D_{exp}^{CO} = 0.82$).

This finding was partially supported by Method two where the control condition EFM weights were best fit by the log-normal distribution. For the Alzheimer's disease EFM weights, the log-normal distribution failed the goodness-of-fit test with Method two indicating the data was best described by a power law distribution. However, it

should be noted that Method two discarded nearly 50% of the total number of EFM weights when fitting the distribution, lending more credibility to results identified from Method one.

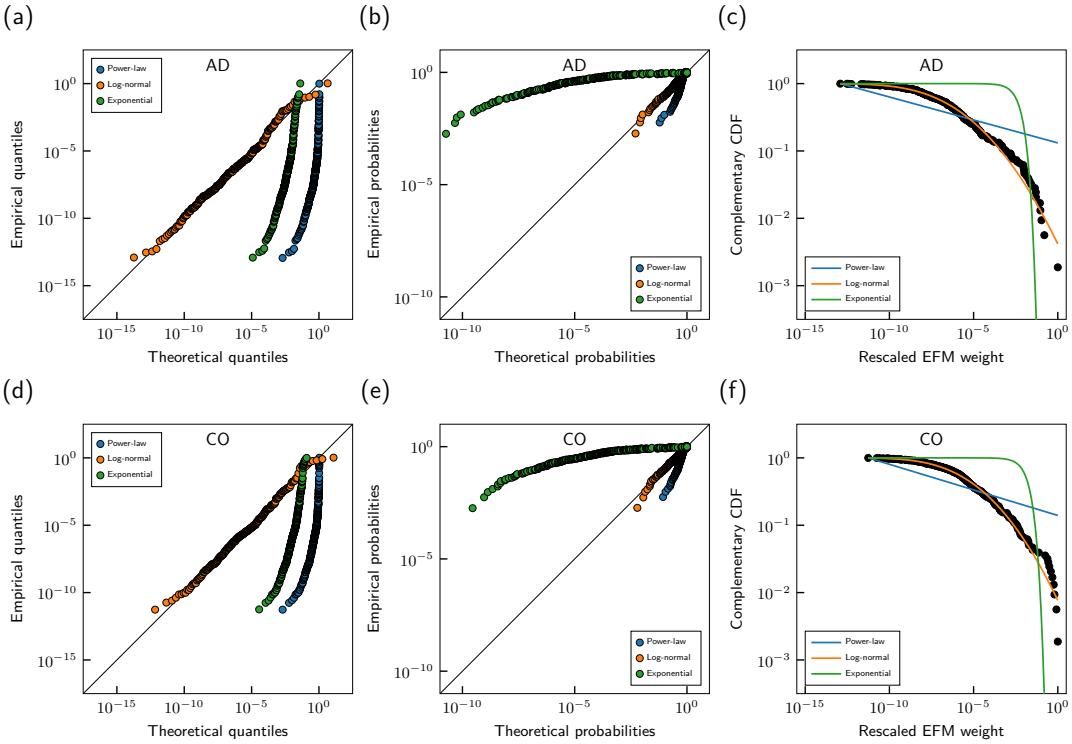


Figure 5.4: Fitted distributions to the rescaled EFM weights within unperturbed conditions using Method one. (a-c) Q-Q plot, P-P plot, and complementary CDF plot of Alzheimer's disease samples. (d-f) Q-Q plot, P-P plot, and complementary CDF plot of control samples.

5.4.3 Analysis of the sample-specific kinetic models

Steady state fluxes approximately follow a log-normal distribution within conditions

I next turned to analysing the sample-specific kinetic models to determine whether the enzyme expression coefficients would alter the distribution of reaction fluxes. Using the same statistical analyses as Section 5.4.2, Figure 5.5 shows the distribution fits across 219 sample-specific Alzheimer's disease and 70 sample-specific control samples

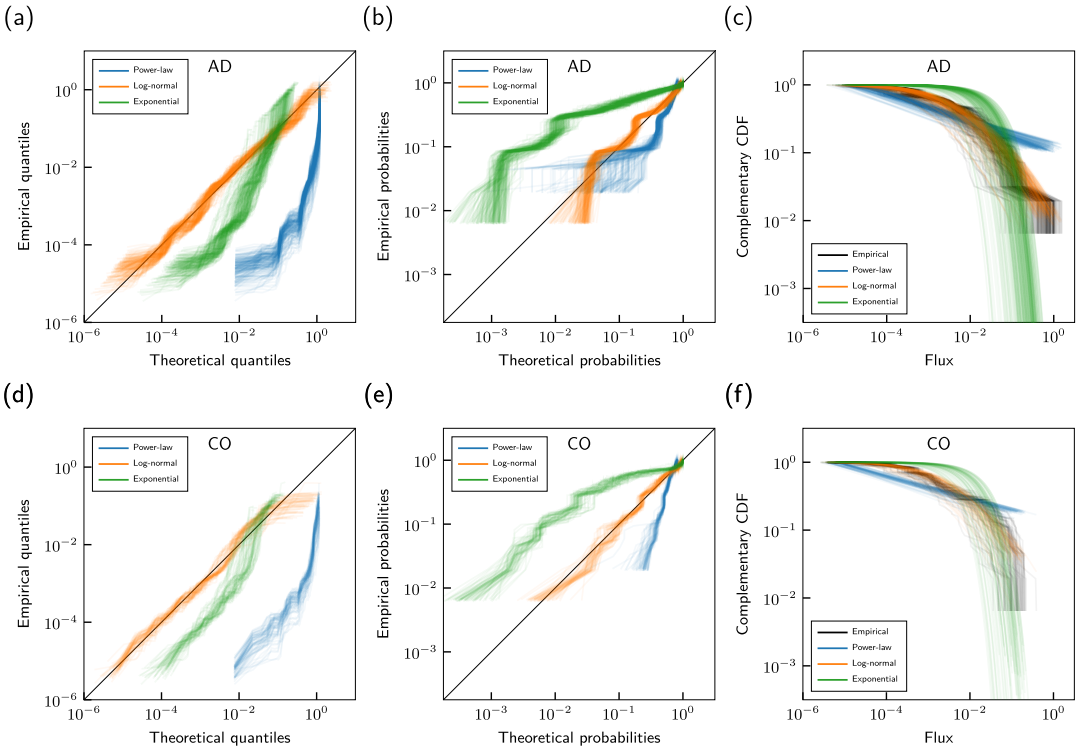


Figure 5.5: Fitted distributions to the reaction fluxes within sample-specific conditions using Method one. (a-c) Q-Q plot, P-P plot, and complementary CDF plot of Alzheimer's disease samples. (d-f) Q-Q plot, P-P plot, and complementary CDF plot of control samples.

using Method one. Following the results from the unperturbed analyses, the Q-Q, P-P, and complementary CDF plots indicated that reaction fluxes remained log-normally distributed in both conditions. This was unanimously confirmed from the goodness-of-fit tests using either the median log-likelihood or KS statistic ($\tilde{D}_{ln}^{AD} = 0.11$, $\tilde{D}_{pl}^{AD} = 0.25$, $\tilde{D}_{exp}^{AD} = 0.43$; and $\tilde{D}_{ln}^{CO} = 0.10$, $\tilde{D}_{pl}^{CO} = 0.33$, $\tilde{D}_{exp}^{CO} = 0.45$). This result is potentially unsurprising since the enzyme expression coefficients only altered $V_{\max}$ by approximately 1.5 orders of magnitude (see Figure 5.2). These perturbations may therefore be insufficient to alter the log-normal distribution of unperturbed fluxes in either condition.

Interpreting the goodness-of-fit using Method two is much harder since excluding data will often improve the fit of any model. A bootstrapped $P > 0.1$ was used to evaluate the plausibility of sample fluxes following a given distribution. Figure B.1a-b shows

that the greatest number of Alzheimer's disease samples were plausibly log-normally distributed while on average maintaining the smallest $x_{\min}$ values. Interestingly, results for the control dataset in Figure B.1c-d suggested that fluxes were exponentially distributed in 59% of samples. However, the exponential distribution also excluded over 30% of the fluxes, on average, in contrast with 36% of samples that were plausibly log-normally distributed and included over 90% of the flux values on average.

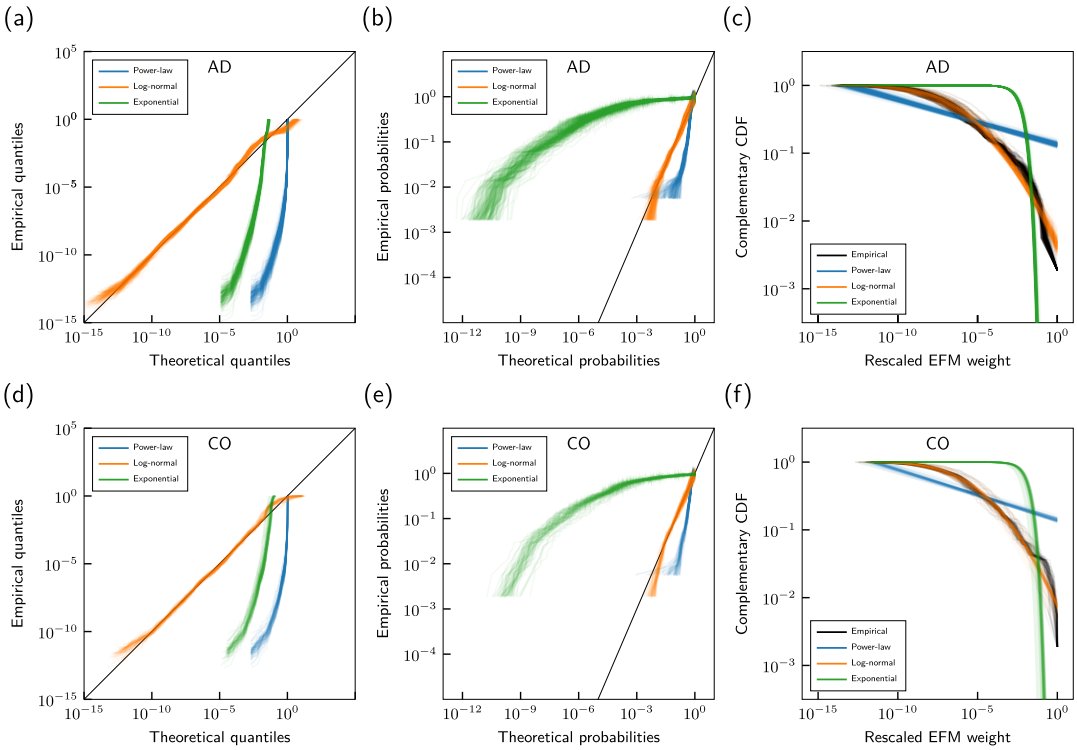


Figure 5.6: Fitted distributions to the rescaled EFM fluxes within sample-specific conditions using Method one. (a-c) Q-Q plot, P-P plot, and complementary CDF plot of Alzheimer's disease samples. (d-f) Q-Q plot, P-P plot, and complementary CDF plot of control samples.

EFM weights approximately follow a log-normal distribution within conditions

Having confirmed that the reaction fluxes within the sample-specific model were approximately log-normally distributed, I then applied the same statistical analyses in Section 5.4.2 to examine the corresponding rescaled EFM weights. Using any goodness-of-

fit test in Method one, Figure 5.6 revealed that pathway flux dominance in each sample was best modelled by a log-normal fit based on the overlaid Q-Q, P-P, and complementary CDF plots within both conditions ($\tilde{D}_{ln}^{AD} = 0.043$, $\tilde{D}_{pl}^{AD} = 0.32$, $\tilde{D}_{exp}^{AD} = 0.84$, $D_{ln}^{CO} = 0.043$, $\tilde{D}_{pl}^{CO} = 0.30$, $\tilde{D}_{exp}^{CO} = 0.82$). These results were also confirmed using Method two based on the rescaled EFM weights (Figure B.2) where the exponential distribution was the most plausible distribution with the smallest $x_{\min}$ to prevent overfitting to a small subset of EFM weights.

Log-normal pathway flux dominance is independent of the distribution of $V_{\max}$ parameters

In hindsight, the log-normal distribution of reaction fluxes and EFM weights revealed in the previous analyses may be explained by the distribution of kinetic parameters in the unperturbed kinetic model. Specifically, a log-normal distribution of $V_{\max}$ parameters would enable reaction fluxes to span several orders of magnitude which could in turn give rise to log-normally distributed pathway fluxes. Using Method one, I indeed found that the distribution of $V_{\max}$ was somewhat approximated by a log-normal distribution spanning 3.5 orders of magnitude (Figure B.3). To test how pathway flux dominance is affected by $V_{\max}$ parameters, I log-uniformly sampled enzyme expression coefficients spanning 5 orders of magnitude (10^{-3} to 10^{2}). From these synthetic perturbations, I then computed the steady state fluxes and corresponding EFM weights for 100,000 simulations based on the kinetic parameters from the control sphingolipid model of Alzheimer's disease.

Since plotting all 100,000 simulation EFM weights would lead to visual oversaturation on a Q-Q, P-P, or CDF plot, I first used Method one to evaluate the goodness-of-fit for the log-normal, power law, and exponential distributions. A median KS statistic of $D_{ln}^{sim} = 0.048$, $D_{pl}^{sim} = 0.32$, and $D_{exp}^{sim} = 0.79$ confirmed that the log-normal distribution exhibited the best fit, especially compared to the other distributions. Comparing the

individual KS statistics within each simulation, I found that $> 99.7\%$ of all simulations were best approximated by the log-normal distribution followed by the exponential distribution for the remaining simulations.

To obtain a broad understanding of the EFM weight distributions, Figure 5.7 shows the densities of the ranked, rescaled EFM weights for the sample-specific Alzheimer's disease, control, and simulated data. The transcript-guided rescaled EFM weights in Figure 5.7a-b are identical to the data plotted in Figure 5.6. But in this new representation, one can observe that the sample-specific EFMs exhibit dominance via the tilde-shaped tails exhibited by the rapidly decreasing weights on the logarithmic scale. From these densities, one can observe that the rescaled EFM weights for the Alzheimer's

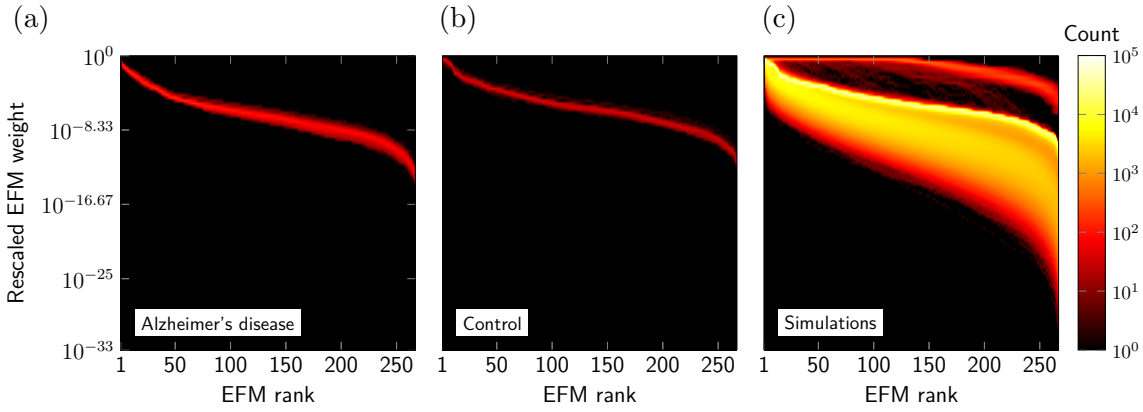


Figure 5.7: Densities of ranked, rescaled EFM weights for the (a) Alzheimer's disease (b) control model, and (c) simulation data.

disease condition decreased more rapidly than the control condition or simulations. This dominance can be quantified by calculating the median rescaled EFM weights associated with a given rank within sample-specific individuals and simulation. Table 5.1 presents these summary statistics where the rescaled EFM weights for Alzheimer's disease are 1-2 orders of magnitude smaller than those computed in the control data. Although the rescaled EFM weights of the simulated data spanned over 30 orders of magnitude, indicating different degrees of EFM dominance, the median values were approximately the same as those calculated for the control samples. Inspecting Figure 5.7c, I found

Table 5.1: Median rescaled EFM weights for selected ranks.

Rank	Alzheimer's disease	Control	Simulations
10	2.38e-02	1.36e-01	1.02e-01
50	2.84e-05	2.62e-04	5.47e-04
150	1.14e-07	1.05e-06	1.99e-06
267	7.35e-14	2.85e-12	9.38e-12

that approximately 800 samples ($< 1\%$) exhibited relatively uniformly distributed EFM weights that did not follow the tilde-shaped tail exhibited in the Alzheimer's disease and control samples. These cases possibly correspond to simulations where the randomly sampled enzyme expression coefficients resulted in perturbed $V_{\max}$ parameters that were approximately uniformly distributed. Overall, these analyses therefore suggest that log-normally distributed EFM weights arise in the sphingolipid kinetic model, regardless of the distribution of $V_{\max}$ parameters.

Specific EFMs tend to dominate across both Alzheimer's disease and control conditions

I next addressed whether particular EFMs were more likely to exhibit dominance over other pathways in both sample-specific sphingolipid kinetic models and the simulation data. By ranking EFMs by their weights, Figure 5.8 shows the distribution of rank counts within the Alzheimer's disease, control, and simulation data. The EFMs on the Y-axis are sorted by highest average rank (rank 1 corresponding to the most dominant EFM), while each row corresponds to the histogram of that EFM rank within samples of that condition. Focusing on the Alzheimer's disease condition, Figure 5.8a shows very narrow rank distributions for the top 50 and bottom 17 EFMs, indicating these pathways are consistently the largest and smallest EFMs across all 217 Alzheimer's disease samples. The remaining EFMs exhibited much more variable ranks spanning nearly a third of the total number of EFMs. The distribution of EFM ranks was much

broader in the control condition, possibly due to the fewer number of samples and greater variability in gene expression coefficients (see Figure 5.2). These data, combined with Figure 5.7a-b, suggest that pathway fluxes are more constrained in the perturbed kinetic models of Alzheimer's disease since flux is being routed through fewer EFMs. This idea of metabolic inflexibility is supported somewhat by the simulation results in Figure 5.8c where the log-uniform perturbations generally result in very narrow rank distributions for the most highly active and inactive EFMs.

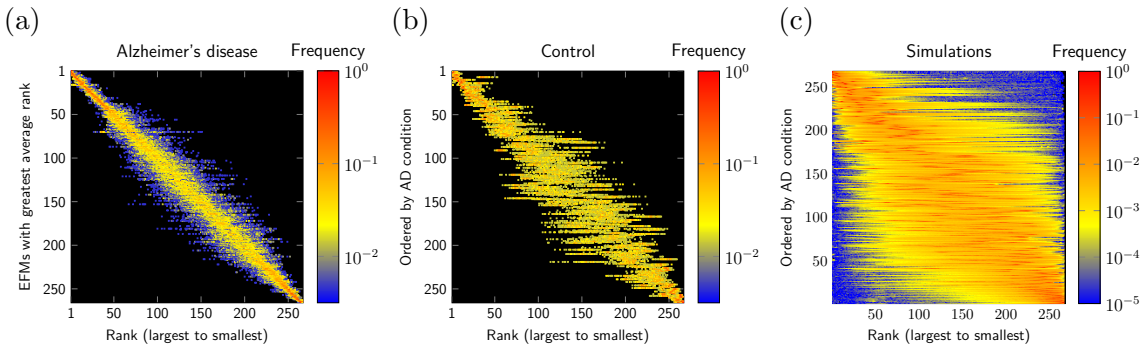


Figure 5.8: 2D histograms of EFM weight ranks for the sample-specific (a) Alzheimer's disease, (b) control model, and (c) simulation data.

5.4.4 Explaining why EFMs are log-normally distributed in both unperturbed and perturbed kinetic models

Thus far, the results have shown that the distribution of reaction fluxes and EFM weights in the unperturbed/sample-specific conditions and simulation data are well-approximated by log-normal distributions. On one hand, dominant reaction fluxes and pathway fluxes may be explained by a log-normal distribution of $V_{\max}$ parameters that naturally lead to log-normal EFM weight distributions. However, the $V_{\max}$ parameters are only very roughly log-normally distributed in Figure B.3 and may follow different distributions, such as the log-uniform range I used in the last simulation. And yet, in all cases, I found compelling log-normally distributed EFM weights.

Here, I demonstrate this formally with a family of metabolic networks that generalize a variant of the metabolic network shown in Figure 3.1. One explanation for these results is that the distribution of EFM weights may arise from the CHMC and how it assigns EFM weights under the Markov constraint. When working with acyclic networks, the CHMC computes source-to-sink EFM probabilities by multiplying the probabilities of individual reactions occurring along a given pathway. An example of this is shown in Figure 5.9 where the corresponding toy network consists of repeating sequences of forked reactions. Each fork carries a steady state flux of α and β before converging at the subsequent node. Each additional repeating sequence doubles the

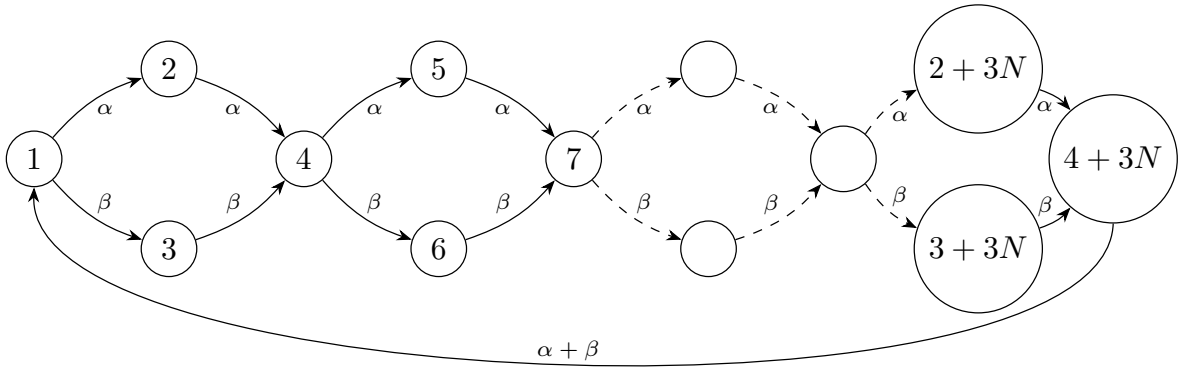


Figure 5.9: Toy network with nodes corresponding to metabolite indices and edges corresponding to metabolic fluxes. The molecule entering the network at node 1 transits through the upper or lower branch with α and β flux, respectively, and continues N times before revisiting node 1.

number of EFMs because all previous EFMs can transit the upper or lower fork of the newly added sequence to maintain steady state for 2^N EFMs. For the purpose of computing EFM probabilities, this network can be viewed as a directed acyclic graph since all EFMs loop back to node 1. Under the CHMC proposed in Chapter 3, each EFM probability of this graph is equal to the product of the individual path probabilities. For the upper fork, this probability is $p_\alpha = \alpha/(\alpha+\beta)$; for the lower fork, this probability is $p_\beta = 1 - p_\alpha$.

Imagine selecting an EFM in this network with K upper forks and $N - K$ lower

forks. Then the weight of that path is

$$w = p_\alpha^K \cdot p_\beta^{(N-K)}. \tag{5.8}$$

I would like to know how these weights are distributed across all possible EFM paths. In particular, I hypothesize that these weights are approximately log-normally distributed. To see this, consider the logarithm of the path weight

$$\begin{aligned} \log(w) &= K\log(p_\alpha) + (N-K)\log(p_\beta) \\ &= K\log\left(\frac{p_\alpha}{p_\beta}\right) + N\log(p_\beta). \end{aligned} \tag{5.9}$$

Observe that $K \sim\ Bin(N, p_\alpha)$ is a binomial random variable with parameter N denoting the number of transited upper forks with probability p_α. The mean and variance of K is

$$E[K] = Np_\alpha \quad \text{and} \quad Var[K] = Np_\alpha \cdot (1-p_\alpha). \tag{5.10}$$

As N increases, it can be shown using de Moivre's theorem[224] that the discrete random variable K can be approximated by the continuous random variable $K' \sim\ \mathcal{N}(Np_\alpha, [Np_\alpha \cdot (1-p_\alpha)]^{1/2})$. Therefore, Equation (5.9) can be approximated as

$$\log(w) \approx K'\log\left(\frac{p_\alpha}{p_\beta}\right) + N\log(p_\beta). \tag{5.11}$$

Observe again that K' is multiplied by the constant $\log(p_\alpha/p_\beta)$ plus the constant $N\log(p_\beta)$. For a normally distributed variable, these multiplication and addition operations do not change the distribution (K' remains a normal random variable) and only alter the mean and variance with

$$\begin{aligned} E[K'\log(\alpha/\beta) + N\log(\beta)] &= Np_\alpha \cdot K'\log(p_\alpha/p_\beta) + N\log(p_\beta) \\ Var[K'\log(\alpha/\beta) + N\log(\beta)] &= Np_\alpha \cdot (1-p_\alpha) \cdot (K'\log(p_\alpha/p_\beta))^2. \end{aligned}$$

Hence, the probabilities assigned by the CHMC for this toy network will indeed approximately follow a log-normal distribution. This observation is true whether the graph contains multiple forks or even varying numbers of forks since transition probabilities along every path are being multiplied together. If I have a series of N forks, each could have different numbers of branches and different probabilities. The path weight would be proportional to the product of whatever is chosen at each branch. The log weights would therefore be the sum of the logs of the corresponding outgoing fractions. In general, the sum of random variables is normally distributed by the central limit theorem and therefore the log weights should also be. That being said, this analysis does not fully account for closed-loop EFMs involving cyclic reactions since their probabilities are not necessarily equal to the individual path probabilities. As discussed in Chapter 3, there may be multiple simple cycles corresponding to these EFMs and their simple cycle probabilities are summed together according to the CHMC algorithm. Nevertheless, this result provides an explanation for the improved log-normal fits observed between the EFM weights compared to the reaction fluxes. It is therefore possible that pathway flux dominance may also follow a log-normal distribution in large-scale networks, especially when there is a greater ratio of fluxes associated with source-to-sink versus closed-loop EFMs.

5.5 Discussion and conclusion

In this chapter, I investigated the distribution of reaction fluxes and pathway fluxes in a transcript-guided sphingolipid kinetic model of metabolism. Using two statistical methods, I found that both unperturbed and perturbed kinetic models exhibited both uneven distributions of reaction fluxes and EFM weights. When perturbation magnitudes were increased to sufficiently change the $V_{\max}$ distributions, I found that EFM weights generally remained uneven, specifically log-normally distributed, and charac-

terized by similar sets of highly dominant EFMs. To better understand why EFM weights followed this log-normal distribution specifically, I examined the distribution of pathway fluxes in a theoretical binomial tree network under the CHMC method.

Although these analyses are specific to the sphingolipid network, there is evidence to suggest that pathway flux dominance exists in other metabolic networks. Many forms of FBA solutions in microbial metabolism distribute fluxes unevenly and often towards maximizing the biomass reaction.[100] However, it remains challenging to determine whether pathway flux dominance arises in GEMs since these networks contain multispecies reactions and the CHMC method is limited to strictly unimolecular reaction networks. I note that a solution to this generalized EFM flux decomposition problem is discussed in Chapter 6 with the question of pathway flux dominance examined in the following Chapter 7.

Large-scale networks may also be more prone to exhibiting dominant fluxes because of their increased complexity constraining the distribution of fluxes.[218] Specific metabolic subpathways are also known to be regulated more tightly than others, such as core carbon metabolism which is central for bioenergetic maintenance.[225] Incidentally, this may explain why FBA programs such as PheFlux have been shown to estimate fluxes more accurately in a large-scale network of *E. coli* metabolism versus central carbon metabolism,[113] based on transcriptomics data, and why the statistical flux model BayFlux has been shown to estimate more narrow posterior flux distributions in large-versus-small scale metabolic networks.[138] Thus, it is quite plausible that even stronger dominant pathway fluxes are observed in larger GEMs.

6

On atomic elementary flux modes in metabolic flux networks under steady state

This chapter consists of work from "Atomic elementary flux modes explain the steady state flow of metabolites in large-scale flux networks" by Justin G. Chitpin and Theodore J. Perkins. This work has been submitted for publication and the corresponding preprint can be found on BiorXiv at `https://www.biorxiv.org/content/10.1101/2024.11.13.623484v1`. The algorithm implementation is available at `https://github.com/jchitpin/MarkovWeightedEFMs.jl`, while all data and source codes to reproduce all figures are available at `https://github.com/jchitpin/reproduce-efm-paper-2024`.

6.1 Author contributions

Justin G. Chitpin developed the methods, conducted the experiments, wrote the software, analyzed the results, and wrote the manuscript with guidance from Theodore J. Perkins.

6.2 Abstract

Steady state fluxes are a measure of cellular activity under metabolic homoeostasis, but understanding how individual substrates are metabolized remains a challenge in large-scale networks. Pathway-based approaches such as elementary flux mode (EFM) analysis are limited to small networks due to the combinatorial explosion of pathways and the ambiguity of decomposing fluxes onto EFMs. Here, I present an alternative approach to explain metabolic fluxes in terms of the steady state flow of their molecular constituents through my proposal of atomic elementary flux modes (AEFMs).

6.3 Introduction

Recent technological and computational advancements are leading to a rise in metabolic flux data.[88,194,226,227] These fluxes quantify the rate of metabolite interconversions within a cell or their movements across subcellular compartments. Metabolic fluxes are studied to understand the dynamics of substrate switching,[228,229] redox balancing,[230,231] overflow metabolism,[232,233] cell differentiation,[234,235] and cellular communities.[236–238] Hence, 'fluxomics'[239,240] is emerging as a field dedicated to quantifying the totality of fluxes within biological systems.

Current experimental and computational methods generally estimate fluxes under metabolic homeostasis. Under this assumption, the production and consumption of

metabolites are balanced over time, and the fluxes are said to be under steady state. Techniques such as stable isotope tracing analysis (SITA) experimentally determine steady state fluxes by culturing cells in labelled substrates and quantifying their incorporation within endogenous metabolites.[87,90] Using this method, fluxes have been measured in major metabolic subsystems such as glycolysis, pentose phosphate pathway, and the tricarboxylic acid (TCA) cycle.[36,241,242] Many computational methods have also been proposed to infer steady state fluxes from transcript, protein, and metabolite abundance data.[111,113,206,243,244] Collectively, these approaches are being used to estimate and predict fluxes in genome-scale metabolic models (GEMs) involving hundreds to thousands of reactions.[79,89,245]

Although much effort has been made to estimate steady state fluxes, less emphasis has been placed on developing methods to analyze these large-scale flux datasets. In contrast to abundance data (e.g. transcripts, proteins, metabolites), steady state fluxes are constrained by the reaction stoichiometries between metabolite inflows and outflows. In a network with fixed source and sink fluxes, an increase in flux along one metabolic pathway must decrease fluxes along one or more other pathways to maintain mass conservation. For simple networks, one may predict how the system would evolve by identifying the up and downstream reactions flanking an enzymatic perturbation. For example, Schwartz and Kanehisa were able to enumerate all glycolytic pathway fluxes in a 12-reaction model of yeast glycolysis.[55] Yet in more complex genome-scale metabolic networks containing thousands of fluxes (e.g. Brunk et al.,[188] Robinson et al.,[246], and H Wang et al.[247]), there exists no hierarchical relationship between reactions. A change in one reaction rate, let alone multiple perturbations, may alter network fluxes beyond neighbouring reactions.[248]

This problem of understanding the flow of metabolites within steady state flux networks has historically been addressed through the concept of elementary flux modes (EFMs). EFMs are a pathway-based approach to analyze metabolic networks and

are defined as minimal sets of biochemical reactions that maintain steady state flux.[37] This minimal property specifies that an EFM cannot be decomposed into two or more smaller pathways carrying steady state flux. An attractive property of EFMs is that any set of steady state fluxes in a network can be explained as a positive, linear combination of EFMs.[210] While EFMs are a mechanistic explanation of how metabolites travel through networks, the number of EFMs scales combinatorially as a function of network size.[64] It remains computationally infeasible to enumerate EFMs in GEMs after over a decade of algorithmic advancements (e.g. Gagneur and Klamt,[58], Terzer and Stelling,[59], and BA Buchner and Zanghellini[61]). The number of EFMs also poses a challenge for explaining network fluxes in terms of their EFMs. When there are more EFMs than linearly independent fluxes, there exist multiple ways to assign EFM weights that reconstruct the observed network fluxes. Consequently, structural analyses of EFMs and downstream analyses of their weights have been limited to small-scale metabolic subnetworks.[55,249–252]

In Chapter 3, I developed a novel solution to the EFM flux decomposition problem for the special case of unimolecular reaction networks under a Markov constraint. I modelled a hypothetical particle moving between metabolites under the assumption that the transition probability of each reaction was proportional to the observed steady state flux. By expanding the notion of Markov chain state to include sequences of metabolites, I proposed an algorithm to efficiently compute simple cycle probabilities corresponding to a given EFM. I called this construct the cycle-history Markov chain (CHMC) and proved that these EFM probabilities could be scaled to weights that fully reconstructed the flux along each reaction[182] (Appendix A.1).

In this chapter, I generalize the CHMC method to operate on any type of metabolic model, including those involving multispecies reactions. I do so by proposing a new type of AEFM which traces the steady state flow of an individual atom within a metabolic network. I refer to these pathways as AEFMs to distinguish them from the (molecular)

EFMs proposed by S Schuster and Hilgetag.[37] While a linear combination of molecular EFMs will reconstruct the overall reaction fluxes, the totality of AEFM weights for a given source metabolite will always reconstruct its mass flow entering the network. I first introduce the framework for enumerating and uniquely identifying AEFM weights. Using a state-of-the-art atom mapping algorithm, I show how to generate atomic transition graphs from which I construct atomic cycle-history Markov chains (ACHMCs) to enumerate and uniquely assign AEFM weights.

6.4 Overview

6.4.1 From molecular to atomic flows in GEMs

I first describe what molecular EFMs look like in metabolic networks and provide intuition why enumerating and decomposing fluxes onto these pathways remains a fundamental problem in literature. Figure 6.1a shows an example molecular EFM involving substrates glucose and ATP which are consumed to produce metabolites G3P, DHAP, and ADP. Multiple source and sink fluxes are required to balance the internal metabolite stoichiometries, leading to several branching and converging paths within the molecular EFM. As the metabolic networks expands to include dozens or more reactions, the corresponding molecular EFMs can grow increasingly complex with numerous source and sink fluxes required to stoichiometrically balance participating metabolites. In particular, there may be multiple sets of reactions involved to balance pairs of energetic molecules (e.g. ATP/ADP), redox equivalents (e.g. NAD^+/glsxtrshortnadh), metabolic byproducts (e.g. water, hydrogen ions), or coenzymes (e.g. CoA). This complexity can lead to an intractable number of pathways that may only differ by reaction subsets required to mass-balance these metabolites. Choosing metabolites to remove may, in part, solve this problem. However, the choice of metabolites to prune from the net-

work is often subjective and not guaranteed to sufficiently reduce the total number of molecular EFMs.

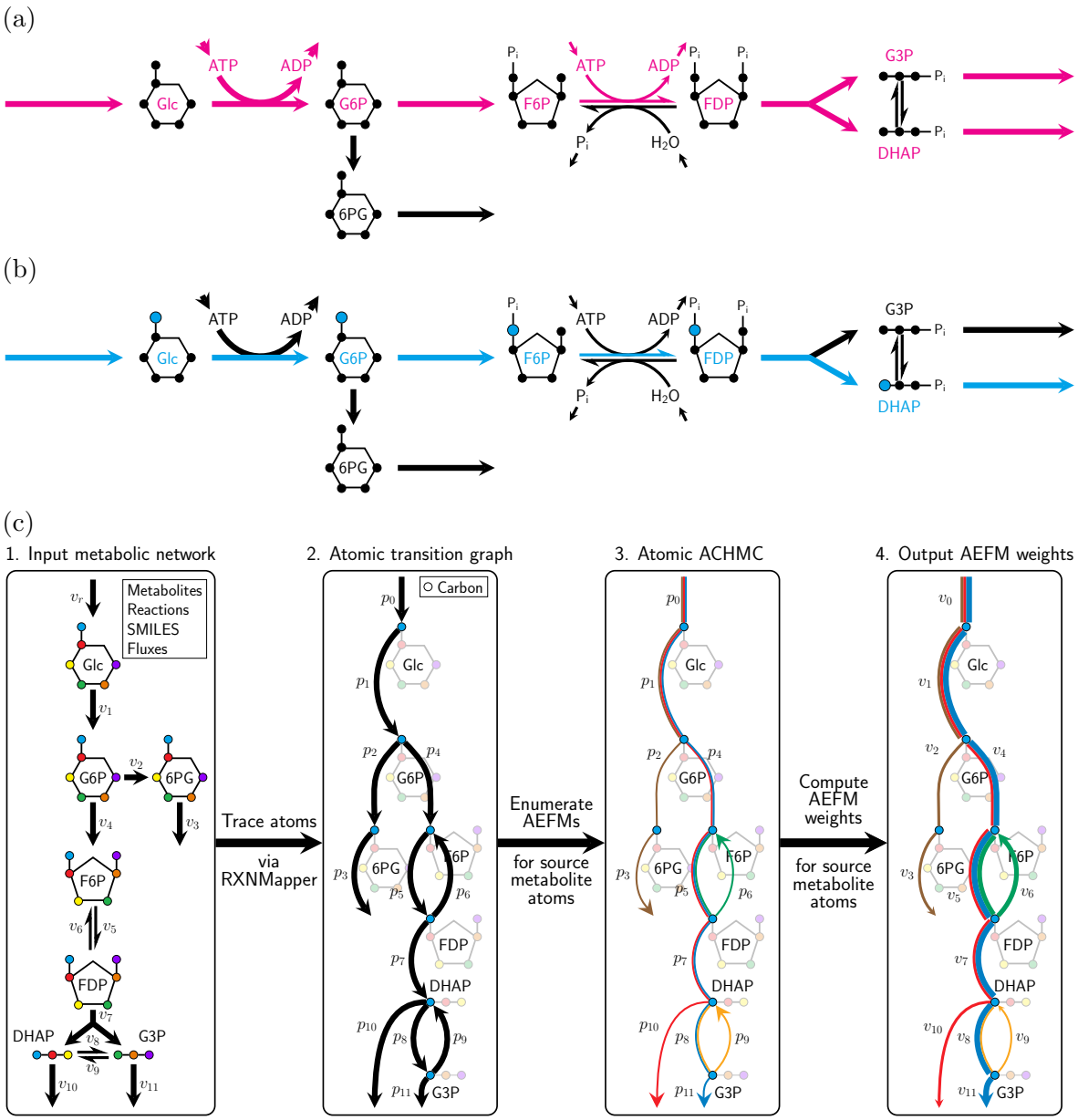


Figure 6.1: Atomic cycle-history Markov chains (ACHMCs) pipeline. (a) An example molecular EFM in glycolysis. (b) An example AEFM tracing the flow of a given glucose carbon in glycolysis. (c) The major steps within my pipeline to enumerate AEFMs and compute their weights.

Given many metabolic flux networks are generated to study nutrient catabolism[253] or engineer desirable metabolic products,[146] I argue for studying source metabolite-

derived AEFMs to reduce the combinatorial number of molecular EFMs. I highlight an example carbon AEFM in Figure 6.1b which shows how the C6 glucose carbon is remodelled during the first steps of glycolysis. As atoms are never split apart or fused together in metabolic reactions, these AEFMs always correspond to simple paths spanning source and sink metabolites or simple cycles between internal metabolites. These pathways are straightforward to interpret because each atom within an AEFM can only transit from a single substrate to product, regardless of the network size, connectivity, or number of participating substrates/products in each reaction. This property further enables us to enumerate and compute the AEFM weights using the previously described CHMC method in Chapter 3 to uniquely decompose steady state fluxes onto EFMs for unimolecular reaction networks.[182]

6.4.2 Pipeline to enumerate and compute AEFM weights

I outline my computational pipeline in Figure 6.1c to enumerate AEFMs and uniquely assign their weights when the steady state network fluxes are known. In step 1, I require as inputs the metabolic network defined by a set of metabolites and reactions encoded as a stoichiometry matrix, the metabolite structures denoted as simplified molecular input entry system (SMILES) strings, and the steady state fluxes. From the stoichiometry matrix and SMILES strings, I construct reaction SMILES strings which are imputed into the atom mapping program RXNMapper,[177] an unsupervised machine learning model, to map the position of each atom within each reaction equation. This atom mapping data is used to construct an atomic transition graph in step 2, which traces the atomic position of an individual atom from a given source metabolite entering the network. I set the transition probabilities proportional to the atomic flux through each reaction, accounting for those with multiple substrate stoichiometries (see Methods). For each atomic transition graph, I construct the corresponding ACHMC to enumerate all AEFMs involving that source metabolite atom in step 3. Finally, the AEFM weights

are computed in step 4 by steady state analysis of the ACHMC. A detailed discussion on the effects of incorrect atom mapping on the ACHMC construction is discussed in chapter 8.

6.5 Methods

6.5.1 Model preliminaries

The structure of a knowledge-driven metabolic flux network is encoded as an m by n stoichiometry matrix S with m distinct metabolites and r irreversible reactions. The fluxes along each biochemical reaction v are positive, real-valued weights that denote the number of times each reaction is occurring per unit time. Under metabolic steady state, the net fluxes consuming and producing each metabolite are balanced and $S \cdot v = 0$. These reactions may be unimolecular–single substrate to single product transformations– or involve multiple substrates, products, or stoichiometric copies of participating metabolites. I do not require the network to be strong connected; however, my method will treat independent components separately. In contrast with my previous unimolecular CHMC model, I impose additional requirements on the input metabolites and reactions. The three-dimensional structure of each metabolite must be known and encoded as a SMILES structure. Where applicable, isomeric SMILES structures are used to avoid ambiguous or incorrect atom mappings. I further require integer-valued stoichiometric coefficients because I consider atoms to be indivisible entities from which I enumerate their AEFMs. I lastly require the input metabolic flux network to be open such that external fluxes enter and leave via source and sink metabolites. This requirement is not a technical limitation of my method but enforced to promote the complete enumeration of all source-to-sink AEFMs remodelling nutrient sources. Overall, these requirements restrict the metabolic flux network from containing any

pseudometabolites or pseudometabolic reactions (e.g. biomass reactions).

A knowledge-driven metabolic network can be uniquely decomposed into a finite set of K molecular EFMs. In a unimolecular reaction network, the metabolites and steady state fluxes can be represented as a graph and the molecular EFMs correspond to weighted simple cycles carrying steady state flux through the metabolite vertices. In a network containing multispecies reactions, the correspondence between molecular EFMs and simple cycles is lost as molecular EFMs may include higher order reactions involving flows across multiple substrates and products. For either type of knowledge-driven metabolic network, their observed steady state fluxes can always be represented as a positive, linear combination of their K molecular EFMs. This flux decomposition problem can be represented by the system of linear equations $Aw = f$, where A is an r by K matrix of zero or natural numbers denoting the number of times each reaction occurs molecular EFM $k = 1, 2, \ldots, K$, where w_k is the molecular EFM weight. When the number of molecular EFMs exceeds the number of linearly independent fluxes, w is underdetermined and there are an infinite number of ways to assign molecular EFM weights to explain the observed, steady state fluxes.

6.5.2 Markovian decomposition of fluxes onto EFM weights

In my previous work[182] (Chapter 3), I proposed Markovian EFM weights as a means to resolve the flux decomposition problem in networks consisting exclusively of unimolecular reactions. I formulated a probabilistic model of single particle flux, in the form of a Markov chain, where I assumed the probability of this hypothetical particle transitioning from one metabolite state to another was proportional to the observed, steady state flux. I further proposed a way of efficiently computing these molecular EFM probabilities and weights by introducing the cycle-history Markov chain (CHMC) which both enumerated the molecular EFMs and uniquely identified their probabilities

which were rescaled to weights based on the absolute flux magnitudes. To generalize this CHMC method to networks containing multispecies reactions, I introduce the concept of AEFMs which can be enumerated and their weights computed from an atomic variant of my CHMC method or atomic cycle-history Markov chain (ACHMC). The following sections provide intuitive justification of my method, while I leave the proof of (atomic/single particle) CHMC algorithm correctness in my original paper[182] (and Appendix A).

Single atom, discrete-time Markov chain

I developed a probabilistic model of atomic flux to compute AEFM weights in steady state metabolic flux networks. At this atomic level, individual atoms within metabolite undergo biochemical reactions and move from the substrate to product side of each reaction equation. As the molecules within metabolic flux network are stoichiometrically balanced under metabolic steady state, so too are the corresponding atomic fluxes.

For a user-specified atom within a given source metabolite, I identify all possible metabolite/atom states that are traversable by that initial state and constrained by the reaction atom mapping predictions. For reactions involving single stoichiometric copies of each substrate and product, the atom of interest will always move from a single substrate to product molecule. If this atom is present in a substrate with multiple stoichiometric copies, I assume there is an equal probability of that atom occurring in either copy. Depending on the substrate stoichiometric copy, that atom may move to (i) different product metabolite/atom states, (ii) the same product but at different atomic positions, or (iii) the same metabolite/atom position. In all cases, I set the atomic flux of the atom transitioning from either stoichiometric copy equal to the reaction fluxes to maintain atomic mass conservation. These atom mapping cases are described in Figure 6.2 with selected biological examples in Figure 6.3.

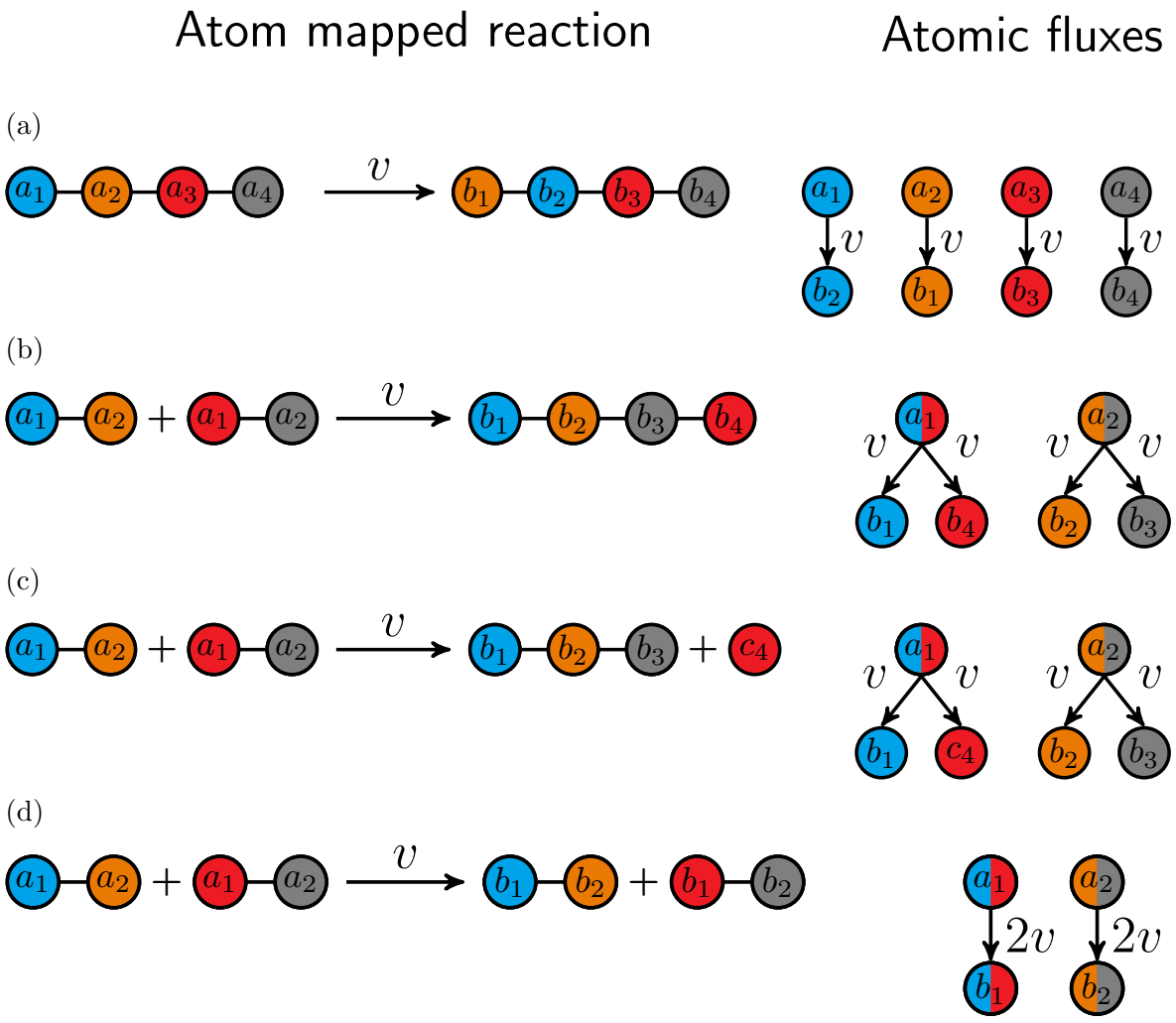


Figure 6.2: Rules for assigning atomic fluxes from atom-mapped reactions. Schematic node colours represent ground truth atom mappings across substrate(s) and product(s). Node labels correspond to distinct atomic indices of substrate and product stoichiometric copies. (a) For reactions with single stoichiometric copies of substrates and products, each atom within a given substrate moves to a single atomic position within a product with atomic flux equal to the reaction flux v. (b-d) For reactions with s stoichiometric copies of a given substrate, there are s copies of each metabolite-atom state on the substrate and product side of the reaction. The probability of any atom within a substrate stoichiometric copy moving to a product metabolite-atom state remains equal to the reaction flux. (b) The stoichiometric copies of the substrate fuse together into a single product. (c) The stoichiometric copies of the substrate form distinct products. (d) The s stoichiometric copies of the substrate form s copies of products with identical atomic backbones.

For a given source metabolite/atom state, the subsequent atomic transitions are encoded as an atomic transition graph $G = (V, E, f)$, where V is the set of nodes corresponding to metabolite/atom states, E is the set of directed edges corresponding

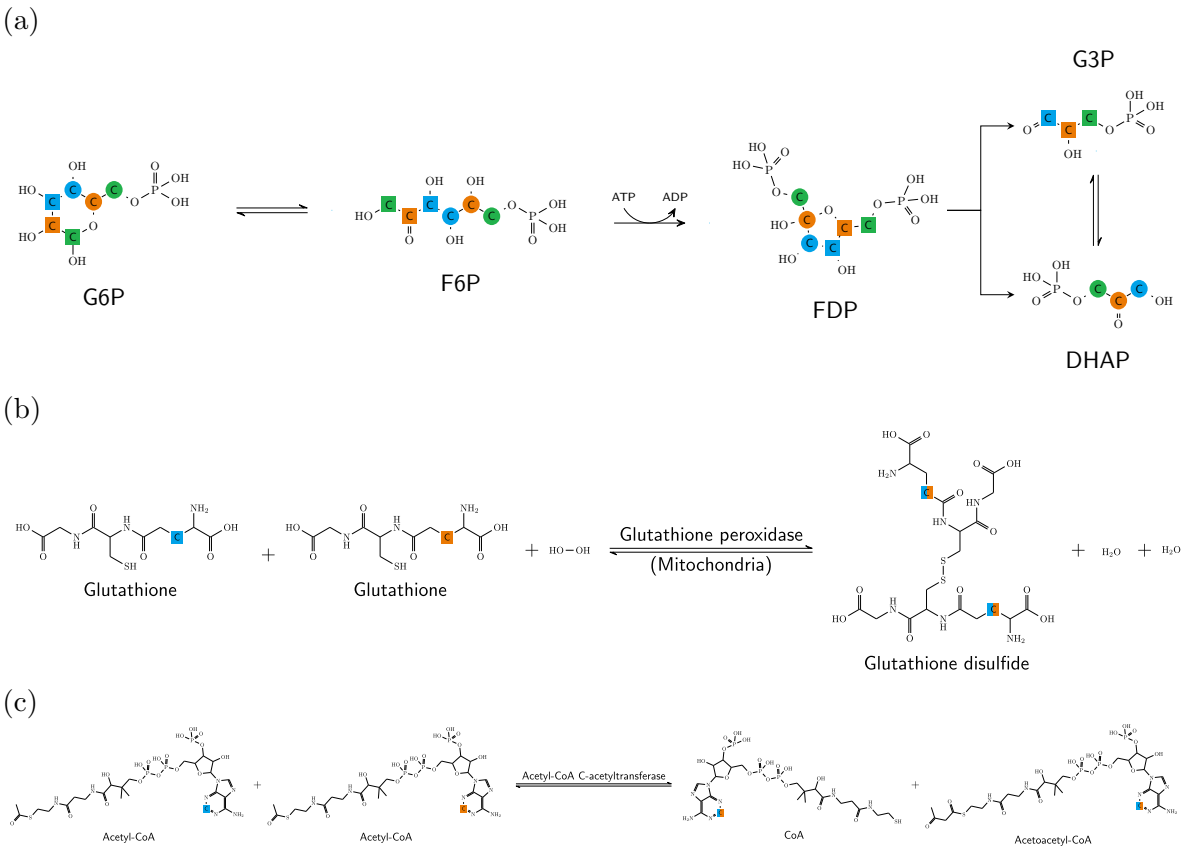


Figure 6.3: Biological examples of specific atom mapping rules. (a) Example of rule in Figure 6.2a where the atoms in each substrate map to a single position on the corresponding product. (b) Example of rule in Figure 6.2b where two stoichiometric copies of glutathione fuse together to form a single stoichiometric copy of glutathione disulfide. (c) Example of rule in Figure 6.2c where two stoichiometric copies of acetyl-CoA undergo a reaction to produce a stoichiometric copy of two distinct products.

to reactions that move the atom from one metabolite to another, and $f \equiv v$ are the atomic mass fluxes. I assume there exists a special node $v_{ext} \in V$ which represents the external environment. Source flux entering the graph occurs along the edge $e = (v_{ext}, r)$, where $r \in V - v_{ext}$ is the source metabolite/atom state. Sink fluxes leaving the graph occur along edges $e = (i, v_{ext})$, where $i \in V - v_{ext}$ are the sink metabolite/atom states. The introduction of v_{ext} closes off the graph and ensures it is strongly connected, where there exists a path in G from any node i to any other node j.

I make a Markovian assumption that the movement of an atom from metabolite/atom position to another in G is proportional to the observed steady state (atomic)

flux of that reaction. I denote these transition probabilities $T_{ij} = f_{ij}/\sum_{j'} f_{ij'}$ for $i \neq j \in |V|$. From here, I follow the same intuition as our (unimolecular) CHMC algorithm[182] to model atomic flows within the network. Starting from an initial source metabolite/atom state s_0, the atom of interest will transit a random sequences of states $s_0, s_1, s_2, \ldots$. As G is strongly connected, that sequence will, with probability of one, include all states infinitely many times. If the atom visits state i, it will eventually return to state i. The sequence of states in between $s_i, \ldots, s_{i'}$ constitutes a cycle, albeit not necessarily a simple cycle. The sequence may include multiple nested or interleaved simple cycles. However, this cycle can always be decomposed into a unique set of simple cycles by repeatedly removing the earliest simple cycle found within the sequence, which I refer to as the "first closure reduction rule." The AEFMs are the set of all simple cycles in G. Thus, I compute the AEFM weights by setting them proportional to the steady state probabilities of their corresponding simple cycles.

Atomic cycle-history Markov chain (ACHMC)

I compute the steady state probabilities of each AEFM following the previously described CHMC method in Chapter 3 for networks of unimolecular reactions.[182] Briefly, the ACHMC is a special type of Markov chain, where I expand the notion of state to include sequences of metabolites transited by the atom. The ACHMC therefore traces the history of a given atom on its way to completing a simple cycle corresponding to an AEFM.

I first initialize the ACHMC on source metabolite atom s_0 within G. Although one could technically choose any arbitrary metabolite/atom state for s_0, it is much easier to interpret AEFMs rooted on a source metabolite flowing into the network. Every state S in the ACHMC corresponds to a simple path $s_0, s_1, s_2, \ldots$. If a reaction transforms s_m to s_{m+1}, the corresponding state $S \equiv s_0, s_1, \ldots, s_m$ transitions to the longer path

$S' \equiv s_0, s_1, \ldots, s_m, s_{m+1}$. If a reaction transforms $s_{m+1} = s_i$, the corresponding state $S \equiv s_0, s_1, \ldots, s_m$ transitions to the shorter path $S' \equiv s_0, s_1, \ldots, s_i$ For either case, the ACHMC transition probability is $\mathcal{T}_{S,S'} = T_{s_m,s_{m+1}}$.

Computing steady state AEFM probabilities

Once constructed, the ACHMC describes all possible simple cycles that a given source metabolite atom can transit through. These simple cycle probabilities are computed by analyzing the steady state dynamics of the ACHMC. As the atomic transition graph G is strongly connected, the corresponding ACHMC will also be strongly connected. There thus exists a stationary distribution π which corresponds to the steady state probability of the atom occupying each ACHMC state.

To compute AEFM probabilities, I first identify all transitions E_K in the ACHMC that close simple cycles corresponding to EFM k. For source-to-sink AEFMs, there exists only a single simple cycle and therefore only one corresponding simple cycle. For internally looped AEFMs, there may exist multiple simple cycles starting at different metabolite/atom states, counting towards the same AEFM. I define the steady state probability of transiting AEFM k as the sum of the steady state probabilities of those transitions or:

$P_k = \sum_{(S,S') \in E_k} \pi_S \cdot \mathcal{T}_{S,S'}$.

As the AEFM probabilities are proportional to the atomic fluxes, these probabilities may be scaled to weights based on the absolute magnitude of the fluxes. Because I enforce that each source metabolite is produced by an input reaction with stoichiometry of one, the sum of all corresponding source-to-sink AEFMs must be proportional to the input metabolite flux. If $l \in K$ denotes all source-to-sink AEFMs flowing through a given source metabolite m_0 and f_0 is the input flux producing that metabolite, I define the proportionality constant $\alpha = f_0 / \sum_l P_l$. This scaling not only ensures the total

input/output fluxes are correctly reconstructed, but also that every individual flux in the network is correctly explained by the totality of AEFM weights.

6.6 Conclusion

By defining EFMs with respect to individual atoms, I described a method for enumerating AEFMs and computing their weights under a Markovian constraint. This workflow leveraged my CHMC method from Chapter 3 and atom mapping program RXNMapper to identify and explain the contribution of individual source metabolite atoms through a metabolic network.

7

Atomic elementary flux mode analysis of five large-scale metabolic networks

This chapter consists of work from "Atomic elementary flux modes explain the steady state flow of metabolites in large-scale flux networks" by Justin G. Chitpin and Theodore J. Perkins. This work has been submitted for publication and the corresponding preprint can be found on BiorXiv at `https://www.biorxiv.org/content/10.1101/2024.11.13.623484v1`. The algorithm implementation is available at `https://github.com/jchitpin/MarkovWeightedEFMs.jl`, while all data and source codes to reproduce all figures are available at `https://github.com/jchitpin/reproduce-efm-paper-2024`.

7.1 Author contributions

Justin G. Chitpin developed the methods, conducted the experiments, wrote the software, analyzed the results, and wrote the manuscript with guidance from Theodore J. Perkins.

7.2 Abstract

I proposed the concept of the atomic elementary flux mode (AEFM) in Chapter 6 to explain the flow of metabolic constituents in steady state flux networks. Here, I show that computations involving AEFMs are orders of magnitude faster than standard (molecular) elementary flux modes (EFMs) which balance molecular stoichiometries. Using this approach, I enumerate carbon and nitrogen AEFMs in five genome-scale metabolic models (GEMs) and compute the AEFM decomposition of fluxes estimated in a human liver cancer cell line (HepG2). These results systematically characterize source metabolite remodelling and, on the HepG2 network, predict glutamine metabolism through the recently discovered non-canonical tricarboxylic acid (TCA) cycle.

7.3 Introduction

In the previous chapter, I introduced the atomic elementary flux mode (AEFM) and described a pipeline to enumerate AEFMs and uniquely identify their weights through my atomic cycle-history Markov chain (ACHMC) method developed in Chapter 6 with the atom mapping program RXNMapper.[177] In this chapter, I follow up on this work by applying my method to five metabolic networks of varying sizes. I first demonstrate that AEFMs can be enumerated in genome-scale metabolic model (GEM) whose molecular EFMs are infeasible by a state-of-the-art enumeration program. I then provide

intuition on the computational feasibility of AEFM-versus-EFM enumeration. While defining EFMs with respect to atoms may appear to increase computational complexity, these atom mapping data constrain the number of carbon and nitrogen AEFMs by several orders of magnitude compared to their molecular EFM counterparts. As such, my ACHMC implementation computes AEFMs several orders of magnitude faster than the molecular EFM enumeration program FluxModeCalculator, with further speedups gained when running my program in parallel. I next perform a structural analysis of these enumerated AEFMs to characterize all theoretical pathways that explain the steady state movement of individual carbons and nitrogens within the network. I further classify these AEFMs into biologically relevant categories and observe remarkable variation in pathway length and the number of pathways transited by atoms within a given source metabolite.

I conclude this chapter with a carbon AEFM weight analysis of a publicly available human liver cancer cell line (HepG2) flux dataset from Nilsson et al.[79] This dataset was chosen for its extensive steady state fluxes estimated from time-series absolute metabolomic measurements. HepG2 cells, like many other cancer cells, are dependent on glutamine for cell survival and proliferation.[254] Understanding why these cells require glutamine, a non-essential amino acid, remains an open question in literature. Using my ACHMC method, I decompose the single set of steady state fluxes in the HepG2 dataset across all source metabolite carbons. Focusing specifically on glutamine-derived AEFMs, I find that the majority of glutamine carbon mass flow is explained by the top five AEFMs across all five glutamine carbons. I find that these AEFMs predict glutamine carbon remodelling through anapleurotic reactions that regenerate TCA cycle intermediates. Interestingly, these AEFMs also correspond with well-known metabolic subsystems and even predict a recently discovered non-canonical TCA cycle pathway[255] identified in non-small cell lung cancer (NSCLC), mouse embryonic, and immortalized mouse myoblast (C2C12) cells.

7.4 Results

7.4.1 Enumerating AEFMs is computationally tractable in large-scale networks

Given the major computational challenge of enumerating molecular EFMs in networks with more than 100 metabolites and reactions, I first sought to test whether it was feasible to enumerate AEFMs in GEMs. I selected five GEM across distinct organisms in ascending order of network size. These networks include a metabolic model of *E. coli* (E. coli core),[36] human red blood cell (iAB RBC 283),[256] *H. pylori* (iIT341),[257] *S. aureus* (iSB619),[66] and HepG2 cells.[79] These models were pre-processed to ensure unambiguous atom mappings across all reactions. Namely, all metabolites with no known structures (pseudometabolites), and pseudoreactions with non-integer stoichiometries were removed from the networks. The resulting metabolic networks ranged between 71-547 metabolites and 144-974 reactions (Table D.1 and Figure D.1). I then attempted to enumerate the molecular EFMs using FluxModeCalculator, which is the fastest molecular EFM enumeration program.[60] Following my pipeline described in Figure 6.1c, I subsequently compiled the metabolite SMILES strings and constructed ACHMCs rooted on each carbon and nitrogen atom across all source metabolites. I focused on these atoms, in particular, for their metabolic significance over chemical elements such as oxygen or phosphorus. The resulting summary statistics describing the GEMs and ACHMCs are presented in Tables 7.1 and 7.2. Across all carbon ACHMCs, I observed a broad increase in the total number of AEFMs as a function of metabolic network size. The one exception to this observation was the HepG2 network which contained fewer metabolites and reactions than iSB619, yet exhibited two orders of magnitude more carbon AEFMs. Inspection of the HepG2 network revealed it only contained 40 source metabolites compared to the 198 source metabolites in the iSB619 network, indicating a greater degree of network connectivity between internal metabolites (Figure D.1).

Table 7.1: Carbon ACHMC summary statistics for five large-scale metabolic networks.

			Carbon ACHMCs				
GEM	Metabolites	Reactions	Sources	Sinks	States	Transitions	AEFMs
E. coli core	71	144	184	255	14659	24355	6519
iAB RBC 283	274	465	954	1166	37122	51220	7958
iIT341	433	759	1486	2134	86342	120545	30633
iSB619	547	974	2188	2986	169877	246712	59444
HepG2	435	545	376	939	9092000	11615891	1377937

Table 7.2: Nitrogen ACHMC summary statistics for five large-scale metabolic networks.

			Nitrogen ACHMCs				
GEM	Metabolites	Reactions	Sources	Sinks	States	Transitions	AEFMs
E. coli core	71	144	30	40	246	448	204
iAB RBC 283	274	465	161	214	6409	9275	2091
iIT341	433	759	314	447	27440	39761	10440
iSB619	547	974	353	516	36661	54165	14214
HepG2	435	545	72	190	19137	27860	5105

Figure 7.1a shows the total number of atomic and molecular EFMs enumerated by my program MarkovWeightedEFMs.jl and FluxModeCalculator on the five GEMs. Using my ACHMC pipeline, I enumerated between $10^3 - 10^6$ carbon and nitrogen AEFMs within each GEM. In contrast, FluxModeCalculator failed to enumerate molecular EFMs in all but the two smallest E. coli core and iAB RBC 283 datasets, returning $10^3 - 10^4$ times more molecular-versus-atomic EFMs.

By focusing on an atom of interest, my ACHMC approach greatly reduced the number of EFMs to biologically relevant pathways involving nutrient source remodelling. This reduction in pathways resulted in AEFM-versus-EFM enumeration completing $10^3 - 10^4$ times faster (Figure 7.1b). While running FluxModeCalculator in parallel did improve run times by 8.2-8.6 times (Figure D.2), it remained over $10^1 - 10^2$ times slower than enumerating their corresponding carbon and nitrogen AEFMs. I further found that AEFM enumeration actually involved fewer metabolite-atom combinations than the total number of metabolites in a given network. Constrained by the atom

mapping predictions, I observed the average carbon and nitrogen ACHMC across all five GEMs only trainsited 0.41% and 1.7% of the total number of metabolite-carbon and metabolite-nitrogen positions, respectively (Figure D.3). While serial run times in Figure 7.1c scaled quadratically ($\text{slope} = 1.90, \text{intercept} = -6.54$), my program and AEFM approach benefits greatly from parallelism. Each ACHMC can be computed independently from one another and my program is also parallelized on the level of individual ACHMCs to improve AEFM enumeration. For the top 28 largest carbon ACHMCs in the HepG2 network, Figure 7.1d shows a median speedup of 7.1 times when using 32 threads.

7.4.2 Most AEFMs are source-to-sink pathways spanning dozens of reactions

I next turned to characterizing the structural properties of all carbon and nitrogen AEFMs across the five GEMs. Molecular EFMs can be divided into pathway that span source and sink metabolites or internal loops within the metabolic network such as pairs of reversible reactions. Both pathways have clear biological interpretations with source-to-sink pathways explaining the steady state flow of source metabolites on their way to exiting the network, and internally looped pathways corresponding to substrate cycles. At an atomic level, these pathways can be subclassified further into AEFMs that transit sets of distinct metabolites or those that revisit the same metabolite(s), albeit in different atomic positions. This notion of a "metabolite revisitation" is similar to an isotopomer in stable isotope tracing.[87] An example source-to-sink AEFM with this property is a glucose-derived carbon AEFM remodelling through two turns of the TCA cycle; on the second turn, the position of the carbon atom shifts within the TCA intermediates, resulting in different isotopomers[258] and therefore a source-to-sink pathway with a metabolite revisitation under my nomenclature.

To obtain a broad understanding of the types of AEFMs within the five GEMs,

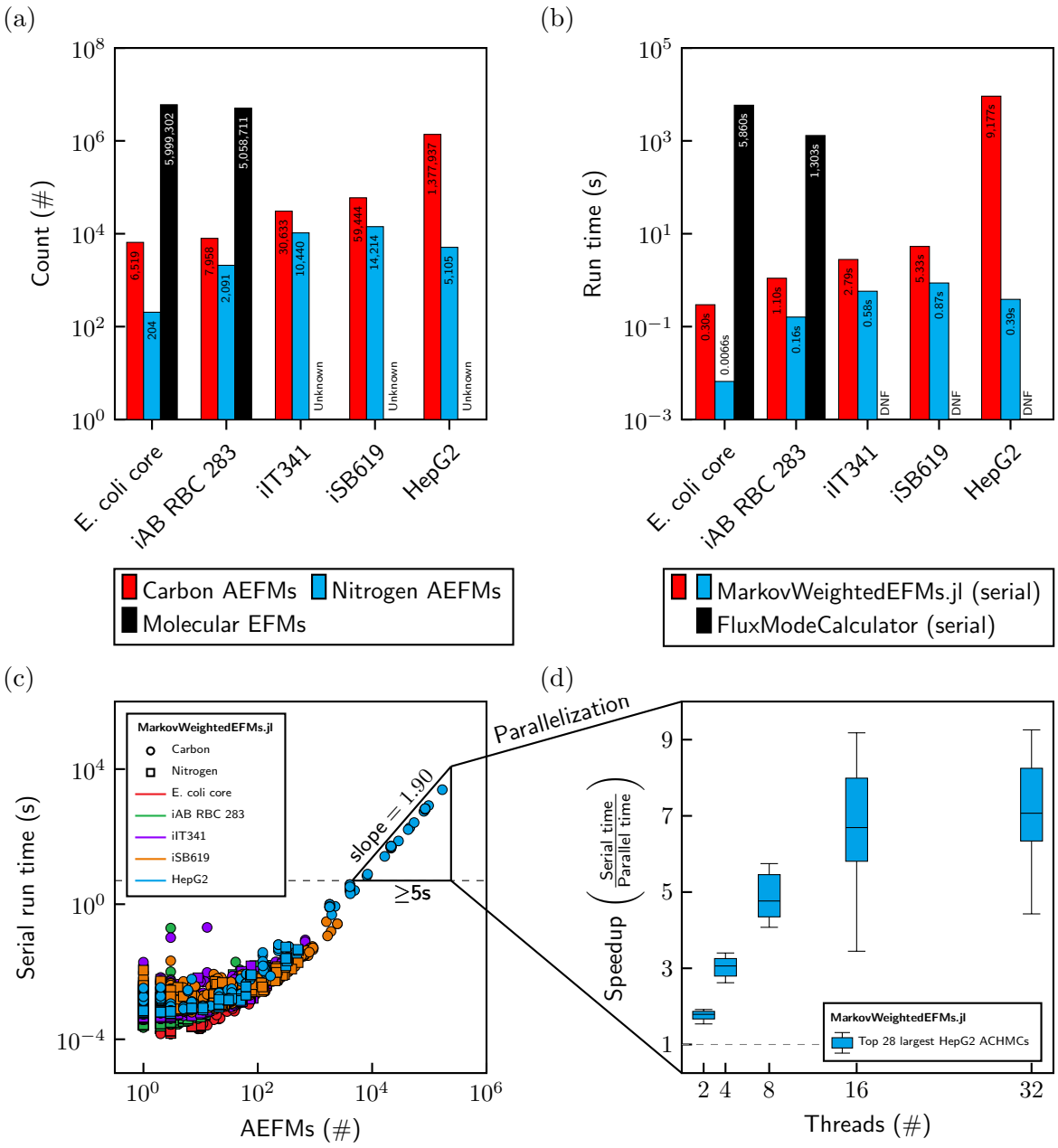


Figure 7.1: AEFMs are computationally tractable compared to molecular EFMs in five GEMs. (a) Number of atomic and molecular EFMs in the five GEMs. (b) The run time of my MarkovWeightedEFMs.jl method versus FluxModeCalculator running in serial. (c) Serial run time scaling of MarkovWeightedEFMs.jl as a function of the number of AEFMs. (D) Speedup of selected ACHMCs as a functional of additional threads.

I categorized them into source-to-sink pathways and looped pathways with or without metabolite revisitations in different atomic positions. Figure 7.2a-b shows the counts for each AEFM classes within each GEM. As expected, I observed that the

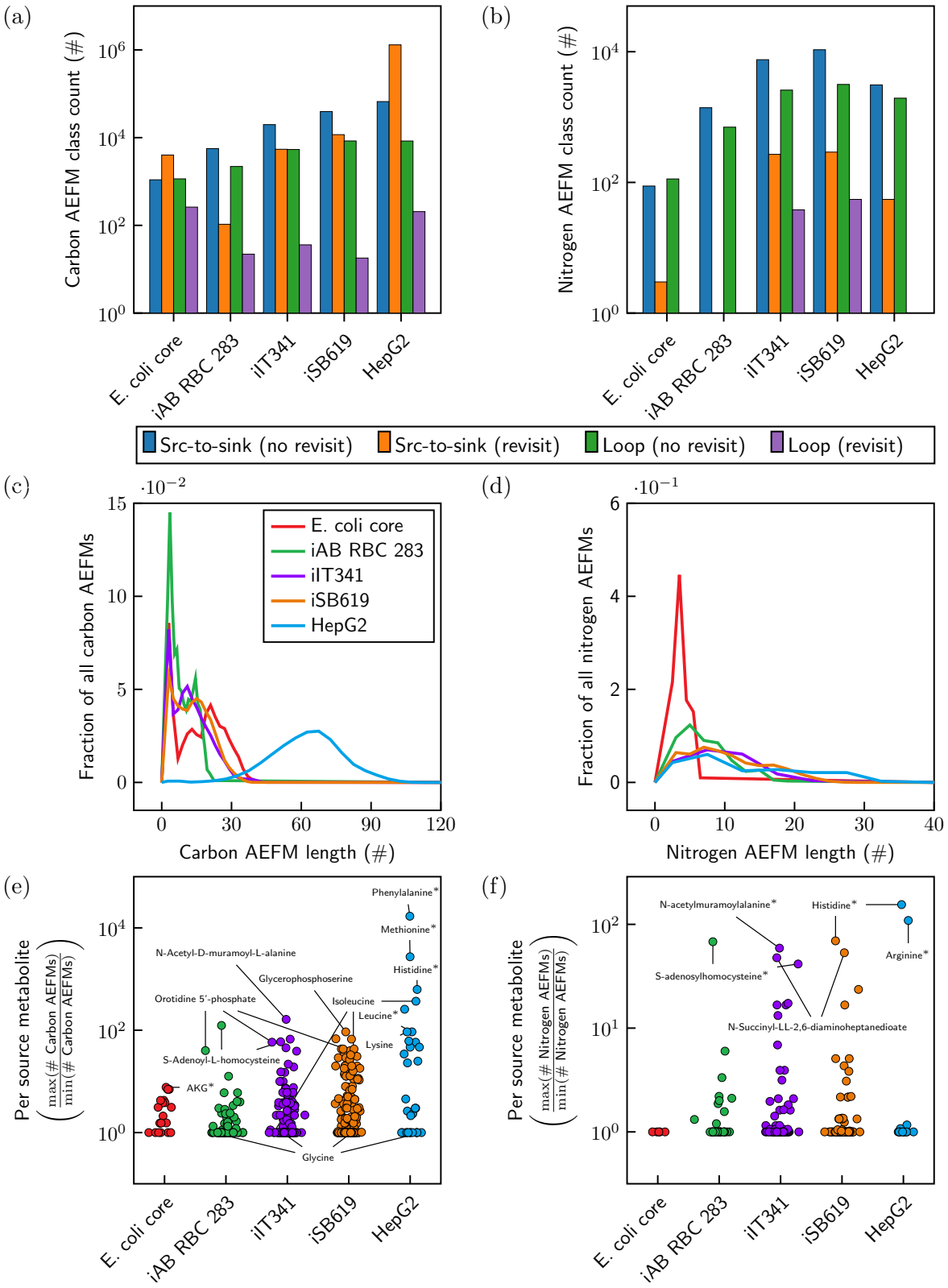


Figure 7.2: Structural analysis of AEFMs in five GEMs.

Figure 7.2 (continued): (a-b) Classes of carbon and nitrogen AEFMs. (c-d) Fraction of carbon and nitrogen AEFM lengths. (e-f) Ratio of the greatest and smallest number of AEFMs between carbons and nitrogens within the same source metabolite. Representative metabolites from each dataset are labelled; labels with asterisks indicate that metabolite is also present in other networks.

majority of carbon AEFMs were source-to-sink pathways with or without metabolite revisitations. Those without metabolite revisitations tended to scale as a function of network size, while those with revisitations increased modestly. Investigation of these source-to-sink pathways confirmed that many metabolite revisitations involved TCA cycle intermediates. Looped AEFMs without metabolite revisitations primarily corresponded to transport reactions cycling metabolites across compartments and cyclic pathways involving TCA intermediates. These two observations explain, in part, why the iSB619 network contained the fewest number of pathways with metabolite revisitations, since they lack organelles and cannot support cyclic AEFMs associated with metabolite transport between compartments nor mitochondrial TCA activity. There were very few internally looped AEFMs with metabolite revisitations with no obvious trend across the five GEMs. The greatest number of these pathways were found in the E. coli core network with many metabolite revisitations involving carbon dioxide derived from isocitrate and alpha-ketoglutarate re-entering the TCA cycle via oxaloacetate. As there are fewer metabolites containing nitrogen atoms, I found the majority of their AEFMs corresponded to pathways without metabolite revisitations, primarily involving amino acid remodelling or transports between compartments. Looped pathways with metabolite revisitations were only observed in the iIT341 and iSB619 networks through reactions involving the production and consumption of ammonia.

Since many looped AEFMs involve reversible reactions or metabolite transport across compartments, I hypothesized that looped AEFMs should involve fewer reactions than source-to-sink pathways which can span numerous internal metabolites. I calculated the fraction of carbon and nitrogen AEFMs with given lengths across all five

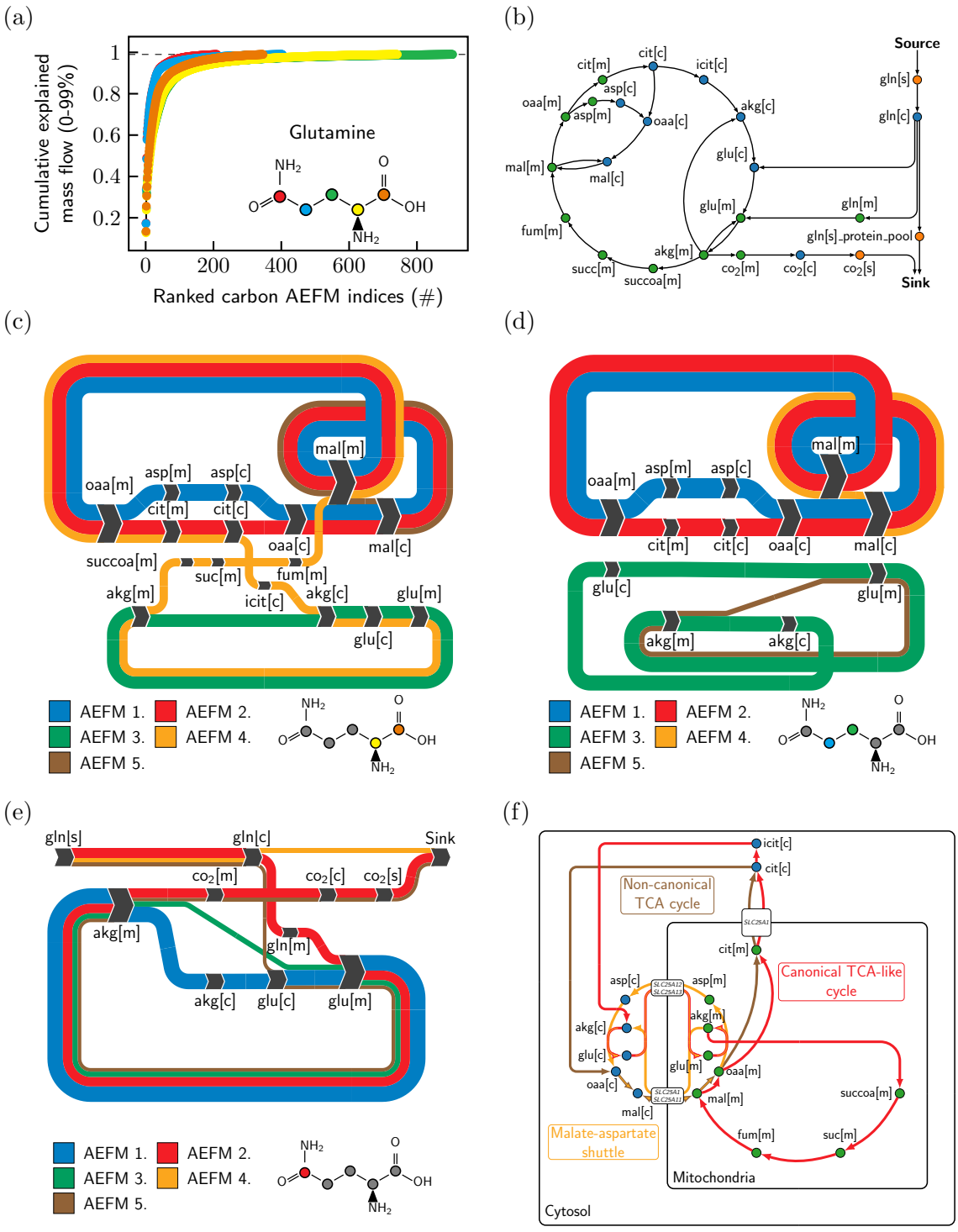


Figure 7.3: Carbon AEFM weight analysis of the HepG2 metabolic flux network.

GEMs and plotted their distributions in Figure 7.2c-d. My results show that carbon AEFMs in the first four networks are bimodally distributed with the majority of path-

Figure 7.3 (continued): (a) A small number of highly active AEFMs explain the majority of glutamine flow. (b) Subnetwork of the top five AEFMs across each glutamine carbon collapsed onto metabolites. (c-e) Sankey diagrams of the top five glutamine carbon AEFMs. (f) Schematic mapping AEFMs onto traditionally defined metabolic pathways/subsystems. All metabolite names are labelled following the BiGG database nomenclature.

ways involving a dozen or fewer reactions. While carbon AEFMs of the HepG2 network appeared to exhibit a single Gaussian distribution ($\mu = 61.7$, $\sigma^2 = 15.5$), closer inspection of the histogram revealed a small peak spanning 2-12 reactions. To determine whether these peaks were associated with source-to-sink and looped AEFM classes, I reclassified the carbon AEFMs spanning less than 5-12 reactions depending on the GEM (shorter pathways), and the remaining carbon AEFMs (longer pathways). I found that longer pathways were associated with source-to-sink pathways, with shorter pathways containing the majority of looped pathways. In fact, 65-95% of all longer pathways in each GEM were source-to-sink pathways with metabolite revisitations, with this statistic increasing to 95-100% when including source-to-sink pathways without metabolite revisitations. Among shorter pathways, I observed that roughly half or more of them were looped carbon AEFMs with or without metabolite revisitations. Analysis of the nitrogen AEFMs revealed only a single distribution of relatively short pathways, since there are much fewer nitrogen-containing metabolites across all GEMs.

I next investigated how atoms within the same source metabolite may follow different paths through the network, irrespective of their AEFM class. I first computed the number of AEFMs for each carbon atom (resp. nitrogen atom). For each metabolite, I then computed the ratio of the class. I then computed the number of AEFMs for each carbon atom (resp. nitrogen atom). So for example, a ratio of 10 means that one of the carbon atoms in the metabolite can transit the network in 10 times as many different ways as the carbon with the fewest paths through the network. My results are shown in Figure 7.2e-f, where each dot corresponds to the ratio for a given

source metabolite. I found that a surprising number of carbons and nitrogens can transit different pathways within their respective source metabolites. Focusing on the source metabolites with carbon and nitrogen ratios over 100, I observed many of these corresponded to hydrophobic and positively-charged amino acids such as phenlylalanine, methionine, arginine, and histidine. Source metabolites containing fewer carbons or nitrogens tended to exhibit ratios closer to 1 with AEFMs involving the same metabolite subsets. Metabolites such as glycine, for example, are equivalently metabolized across the same carbon AEFMs. For each GEM, and not counting source metabolites with ratios of one, I computed median carbon ratios between 2.0 and 46.8 and nitrogen ratios between 1.3 and 11.2. Overall, these findings highlight the varying degrees of metabolic flexibility for atomic constituents of source metabolites to transit the metabolic network.

7.4.3 A minority of carbon AEFMs explains the majority of source metabolite remodelling

I finally highlight the potential of my AEFM method for quantifying nutrient source metabolism. I turned to the HepG2 network which contained a single set of steady state fluxes estimated by Nilsson et al..[79] These fluxes were estimated from a parsimonious flux balance analysis (FBA) model of HepG2 cells cultured under exponential growth with external fluxes constrained by time-resolved exometabolomic measurements (see Methods for more information). My focus was to analyze this flux network using my AEFM framework to identify alternative pathways of glutamine metabolism since many cancer cells, including HepG2 cells, are dependent on this non-essential amino acid for survival and proliferation.[254,259] Using my method, I decomposed the set of steady state fluxes onto the 241,085 glutamine carbon AEFMs characterizing the HepG2 metabolic network. Figure 7.3a shows the cumulative explained mass flow of each glutamine carbon as a function of their AEFMs ranked from greatest to smallest explained mass flow. I capped the cumulative explained mass flow of each glutamine carbon at 99% because I

observed that 98.9% of AEFMs collectively explained the remaining 1% glutamine carbon flow. These results suggested that the majority of glutamine carbon remodelling is explained by a surprisingly small number of glutamine AEFMs, especially since my Markovian flux decomposition approach assigns some weight to all network pathways. The top 5 ranked AEFMs across each glutamine carbon (25 pathways altogether) explained over half (50.2%) of the total glutamine carbon mass flow with the subsequent 5 next highest ranked AEFMs only explaining an additional 9.8% glutamine carbon mass flow. These observations were consistent across carbon AEFMs derived from other amino acids as well as glucose, supporting my dominant pathway flux hypothesis in Chapter 5 (Figure D.4). For some source metabolites, I observed overlapping curves between carbons (e.g. arginine, glycine, serine in Figure D.4), suggesting that individual carbons were remodelled through the same AEFMs. However, corroborating results from Figure 6.2e-f, I observed many non-overlapping curves for the majority of amino acids since their carbons could transit the network via different internal metabolites.

To obtain a high-level understanding of the dominant pathways of glutamine carbon remodelling, I constructed a metabolic subnetwork from the top 5 glutamine carbon AEFMs, collapsing metabolite-atom positions onto distinct metabolites, since none of these AEFMs involved metabolite revisitations. Of the 435 metabolites in the HepG2 dataset, input glutamine can remodel into 202 species localized to the cytosol or mitochondria. I found that the top 5 glutamine-derived carbon AEFMs consist of just 23 metabolites and 31 reactions (Figure 7.3b). These AEFMs generally consisted of cyclic pathways involving TCA intermediates. Only two source-to-sink reactions convert cytosolic glutamine towards the protein pool and carbon dioxide through decarboxylation of alpha-ketoglutarate in the mitochondria.

As AEFMs are steady state pathways, they can be visualized as Sankey diagrams to better understand the carbon flows through the network. Figure 7.3c-e shows the top 5 AEFMs for each glutamine carbon. Each AEFMs in the diagram is shown as a

flow proportional to its AEFM weight with nodes representing the metabolite transited by the given glutamine carbon. Although each glutamine carbon can transit a different number of AEFMs, I found the top 5 backbone carbons and first two sidechain carbons were identical in both AEFM pathway and weight, with no AEFM containing a metabolite revisitation.

Focusing on either backbone carbon in Figure 7.3c, I observed that all AEFMs corresponded to well-established metabolic subsystems. The first, third, and fifth highest ranked AEFMs corresponded to metabolites and reactions involved in the malate-aspartate shuttle. The fourth largest AEFM resembled a combination of canonical TCA cycle activity and malate-aspartate shuttle, with TCA intermediates citrate, isocitrate, and alpha-ketoglutarate localized to the cytosol rather than mitochondria. This AEFM also involved additional reactions transforming alpha-ketoglutarate into glutamate before remodelling into succinyl-CoA in the mitochondria. I noticed that the second largest AEFM cycled carbon mass between TCA intermediates citrate, oxaloacetate, and malate between the cytosol and mitochondria. Interestingly, this metabolic pathway corresponds to the recently discovered non-canonical TCA activity described by Arnold et al., who validated this pathway in NSCLC, mouse embryonic, and C2C12 cells by stable isotope tracing analysis.[255] This non-canonical TCA activity was further observed in glutamine carbons 3 and 4 in AEFM 2 (Figure 7.3d). I also found AEFMs 1 and 3 in Figure 7.3d-e corresponded to the malate-asparate shuttle. Only the terminal glutamine carbon in Figure 7.3e contained source-to-sink AEFMs that decarboxylated glutamine-derived alpha-ketoglutarate into carbon dioxide.

My analyses revealed that the most active glutamine-derived carbon AEFM weights corresponded to well-known metabolic subsystems. This finding led me to ask whether I could summarize my AEFM analysis as a simplified metabolic model describing the majority of glutamine carbon remodelling in the HepG2 flux network. The resulting schematic is shown in Figure 7.3f which consists of the following three ma-

jor metabolic pathways: canonical TCA-like cycle, non-canonical TCA cycle, and the malate-aspartate shuttle. These metabolic pathways were chosen based on the number of top glutamine-derived carbon AEFMs that overlapped with reactions within these subsystems. My results suggest input glutamine is converted into cytosolic glutamate which is transported into the mitochondria through aspartate/glutamate antiporters encoded by solute carrier family 25 member 12 (*SLC25A12*) and solute carrier family 25 member 13 (*SLC25A13*) annotated in the HepG2 metabolic model. Mitochondrial glutamate is then converted into TCA intermediate alpha-ketoglutarate and citrate. Instead of following the canonical reactions of the TCA cycle, the mitochondrial citrate is exported to the cytosol where it can participate in non-canonical TCA cycle activity or the malate-aspartate shuttle to re-enter the mitochondria. Taken together, the dominant carbon AEFMs predicted by my model suggest HepG2 glutamine is conserved within TCA intermediates and regenerated through subsequent reactions involving cytosolic citrate.

7.5 Discussion

Since the proposal of molecular EFMs, much effort has been made to enumerate and decompose steady state metabolic fluxes onto these pathways. By defining EFMs with respect to individual atoms rather than molecules, I introduced the concept of AEFMs which describe how individual atoms are remodelled within a metabolic network. As AEFMs never involve branching pathways, they are easier to interpret than their molecular EFM counterparts which may involve multiple inputs and outputs to balance molecular stoichiometries. Using the atom mapping program RXNMapper, I adapted my previously described cycle-history Markov chain (CHMC) method to efficiently enumerate carbon and nitrogen AEFMs in five GEMs. My structural analysis systematically characterized source metabolite remodelling at the atomic level. By comparing

the number of AEFMs across atoms within the same source metabolite, I revealed how certain amino acids displayed remarkable variability in the number of pathways transited by their constituent atoms. This analysis, in particular, may be useful for designing non-uniformly labelled stable isotope tracers to identify fluxes within specific metabolic subsystems. The AEFM weight analysis of the HepG2 network revealed that the majority of glutamine carbons were remodelled via cyclic pathways regenerating TCA intermediates through cytosolic citrate. Excitingly, my model also predicted non-canonical TCA activity which has not been previously described in HepG2 cells.

As a systematic method to explain the flow of metabolites, the problem of molecular EFM enumeration has historically limited their applicability to small-scale networks. Many algorithmic advances enumeration have incrementally improved molecular EFM enumeration, yet never feasibly beyond networks with > 100 metabolites and reactions.[59,60,260] Molecular EFM enumeration is generally regarded as unfeasible for large-scale networks—so much so that numerous methods have been developed to enumerate a subset of molecular EFMs constrained by pathway length,[261] reaction thermodynamics,[262] participating metabolites,[263] participating reactions,[264] gene regulatory rules,[265] metabolic subsystems,[266,267] extracellular flux measurements,[268,269] or even random sampling.[270] While the HepG2 results did show the majority of source metabolite fluxes were explained by a small number of AEFMs, contextualizing the significance of these dominant pathway fluxes is only possible when all AEFMs are known. Furthermore, reaction thermodynamics, gene regulatory rules, and other constraints may change across fluxes estimated from different biological conditions. These results suggest that enumerating the shortest pathways would fail to recover the majority of TCA pathways, as the top five ranked AEFMs across all glutamine carbons averaged 6 reactions in length. By computing AEFMs, instead, I could feasibly enumerate atomic pathways in genome-scale networks containing hundreds of metabolites and reactions within minutes or hours on desktop computing resources. I am hopeful that algorith-

mic tricks borrowed from molecular EFM enumeration tools, such as linear pathway compression, could further decrease memory usage and run times.

Aside from carbon and nitrogen AEFMs, my work is broadly applicable to other types of chemical elements. One could feasibly enumerate oxygen or hydrogen EFMs in small-scale networks, although the significance of these pathways is unclear for biochemical networks given the ubiquitous nature of these atoms. I also caution against enumerating hydrogen AEFMs, in particular, given limitations associated with RXNMapper. Unless explicitly represented in the reaction SMILES string, RXNMapper does not assign hydrogen atom mappings. While this may seem trivial, RXNMapper cannot map atoms in reaction SMILES strings longer than 512 characters. My method could also be applied to study sulphur metabolism to quantify methionine or cysteine remodelling, or phosphorus metabolism to quantify nucleotide synthesis. Outside the realm of biochemistry, this method is applicable to any periodic table element and possibly relevant for studying general chemical reaction networks.

7.6 Conclusion

My AEFM framework opens many new computational problems in metabolism. While I analyzed one set of fluxes for the HepG2 network, one could envision a differential AEFM analysis of flux distributions under differential conditions. These differential AEFMs may reveal changes in metabolic subsystem activity or predict new pathways of source metabolite remodelling. For instance, analyzing carbon AEFMs of one carbon (C1) compounds may be particularly useful for quantifying or optimizing carbon fixation pathways in C1-utilizing organisms. Another important question is whether joint analysis of AEFM weights could reveal broader metabolic changes within a condition. I observed that many AEFMs across distinct source metabolites converged on common internal metabolites. Intuitively, the weights of AEFMs rooted on different

source metabolites, but sharing common downstream metabolite subsequences, should be related, and this relationship may further speed AEFM enumeration or weight computations. Finally, more work is needed to establish relationships between AEFM decompositions and more traditional EFM decompositions of fluxes, and to explore the possibility that feasible AEFM calculations might somehow aid in EFM calculations.

7.7 Methods

7.7.1 Datasets

Five curated GEMs were chosen in this study to reflect a diversity of organisms, metabolites and biochemical reactions. Four networks were obtained from the Biochemical Genetic and Genomic (BiGG) database[1] in ascending order of network size. Datasets were downloaded as Systems Biology Markup Language (SBML) (.xml) files, and the corresponding stoichiometry matrices, metabolites and reactions were extracted using the Julia SBML.jl package. The HepG2 cell line dataset was chosen for its single set of estimated steady state fluxes estimated from a combination of protein and metabolite abundance data. These cells were grown in Dulbecco's Modified Eagle Medium (DMEM) with 10% fetal calf serum, 22mM glucose, and 1.8mM glutamine.[79] The steady state fluxes were recomputed from the FBA model provided by the authors (figure3_metabolitebalances.m; MATLAB R2018a[79]). Across all models, I modelled all reversible reactions as net forward reactions to simplify the number of molecular and AEFMs as the majority of experimental and computational methods can only estimate net fluxes. I assumed the reaction stoichiometries of the BiGG datasets represented the canonical net directions of all reversible reaction. For the HepG2 dataset, the net directions were chosen based on the net fluxes estimated from theFBA model.

7.7.2 Extracting SMILES strings

SMILES structures were not encoded in the BiGG SBML files nor the HepG2 dataset. I manually extracted simplified molecular input entry system (SMILES) strings corresponding to each metabolite across all GEMs. Isomeric SMILES strings were taken from PubChem,[178] the Human Metabolome Database,[271] or MetaNetX database.[272] Metabolites shared across GEMs were assigned the same SMILES structure in addition to identical metabolites stratified across distinct subcellular compartments. Generic structures such as R-groups were not modelled in this study and were replaced with a single bonded hydrogen under the assumption this modification would not alter atom mappings, especially since RXNMapper does not map hydrogen atoms unless they are explicitly represented in the SMILES strings. SMILES strings were also not assigned to metabolites with ambiguous structures nor pseudometabolites with no known chemical structure. All pseudometabolites and reactions involving pseudometabolites were removed from the stoichiometry matrix (Table D.1).

7.7.3 Constructing reaction SMILES strings

RXNMapper performs atom mapping on a string that encodes the SMILES structures of all substrates and products for a given reaction. These reaction SMILES strings are constructed by concatenating SMILES strings of individual substrates and products. Multiple substrates or products are separated by the symbol '.', while the reaction arrow delimiting substrates and products is denoted by the symbol '>>'. The substrate and product SMILES strings are concatenated in order of metabolite appearance within the stoichiometry matrix, and multiple stoichiometric copies are consecutively repeated in the reaction SMILES strings. Non-integer stoichiometric coefficients are therefore not allowed and these pseudoreactions were removed from the stoichiometry matrix (Table D.1). I manually checked several reactions to ensure the assigned atom mappings

were correct. Crucially, I found RXNMapper to fail on nearly all classes of ATP-independent transaminase reactions. This problem has previously been reported in an atom mapping benchmark study prior to the publication of RXNMapper.[273] These transaminase reactions were identified in all metabolic networks and manually atom mapped based on the KEGG reaction mappings (Table D.1 and Figure D.5).

7.7.4 Pre-processing GEMs

Several pre-processing steps were required to compute AEFMs from the GEMs. The stoichiometric coefficients of external reactions were set to one and their fluxes rescaled to carry equivalent units of source and sink metabolites. All reactions containing pseudometabolites with no known SMILES strings were removed from the stoichiometry matrix and replaced with unimolecular source and sink reactions carrying equivalent units of flux. The same procedure was also applied to reactions containing non-integer stoichiometries. These pseudometabolites were next removed from the stoichiometry matrix rows, and unimolecular reactions involving the same external or internal metabolites were aggregated. Finally, I had to remove a small number of reactions whose reaction SMILES strings exceeded the 512 token limit of RXNMapper (Table D.1). These reactions were likewise converted into unimolecular source and sink reactions to maintain steady state flux across the network. Note that all pre-processing steps described above are wrapped into convenience functions within my Julia package MarkovWeightedEFMs.jl.

7.7.5 Benchmarking enumeration of atomic-versus-standard EFMs

I benchmarked my Julia package MarkovWeightedEFMs.jl version 2.0.2 running Julia version 1.10.2 to enumerate AEFMs versus the molecular EFM enumeration program

FluxModeCalculator[60] running MATLAB version R2018a with default parameters including network compression. I chose FluxModeCalculator because it is the state-of-the-art program for enumerating molecular EFMs. Benchmarks were performed on a 64-bit Linux operating system with 64GB of memory and an AMD 5950X CPU with 16 cores/32 threads. All benchmarks reported the wall time to run the atomic or molecular EFM enumeration functions and did not include startup times to open Julia/MATLAB, precompile Julia functions, or load packages and input data. A benchmark time of did not finish (DNF) was assigned if either program could not complete EFM enumeration within 7 days. In practice, only FluxModeCalculator failed to enumerate the molecular EFMs in three of the five datasets.

ACHMC model

For each carbon and nitrogen atom within a given source metabolite, I identify all possible metabolite/atom states that are traversable by that initial state and constrained by the reaction atom mapping predictions . Following Chitpin and Perkins,[182] these atomic transition graphs were analyzed by an atomic version of my CHMC (ACHMC) method to enumerate the corresponding AEFMs and compute their weights. In this formulation, each ACHMC state corresponds to a sequence of metabolite-atom positions transited by a given source metabolite atom. Each simple cycled identified by the ACHMC corresponds to an AEFM rooted on a user-specified source metabolite atom. For the HepG2 flux network, I computed the probabilities of each AEFM by steady state analysis of the ACHMC. These AEFM probabilities were scaled to weights such that the totality of source-to-sink AEFMs was equal to the unimolecular input flux producing that source metabolite.

8

Discussion

8.1 Summary

As high-throughput methods continuously improve to estimate metabolic fluxes, I anticipate new computational tools will be required to analyze this deluge of data. These tools have historically approached flux analysis via mathematical modelling due to well-established metabolic networks constructed from decades of biochemical studies. Perhaps the most well-known approach has been the elementary flux mode (EFM). Indeed, S Schuster and Hilgetag's seminal EFM paper[37] was the impetus for many subsequent stoichiometric pathway analysis methods.[147,159,161,170] Yet since its inception, EFM analysis has been limited by the fundamental computational challenge of enumerating and uniquely assigning weights to these pathways.

In this thesis, I carefully re-examined the underlying challenges associated with EFM enumeration and analysis and developed novel solutions to overcome these prob-

lems. This journey began in Chapter 3 where I explained how existing methods relying on mathematical optimization could fail to uniquely decompose fluxes onto EFMs in strictly unimolecular reaction networks. I demonstrated these issues on simple toy networks where EFM weights differed according to the chosen objective function and mathematical optimization solver. This finding motivated a more biophysically natural approach to predicting pathway fluxes. As my first major computational contribution to the EFM field, I introduced a probabilistic model of single-particle flux and the cycle-history Markov chain (CHMC) algorithm for efficiently enumerating EFMs and uniquely computing their weights under a Markov constraint. Subsequent work in Chapter 4 demonstrated that this Markovian approach was equally applicable to kinetic models of flux, yielding identical EFM probabilities from kinetic parameters alone. The CHMC method was extensively used in Chapter 5 to analyse a sphinogolipid kinetic model that incorporated RNA-sequencing data to modulate the reaction rates. Analysis of these weights revealed that the majority of pathway fluxes were accounted for by a small number of EFMs, leading me to formulate a *dominant pathway flux hypothesis* of metabolic networks. While this work was limited to unimolecular reaction networks, these techniques and analyses have set the mathematical framework and motivation to generalize and apply my method to larger, multispecies reaction networks.

In the second half of my thesis, I addressed how my CHMC method could be applied to large-scale metabolic networks not limited to strictly unimolecular reactions. By examining the computational and biological shortcomings of (molecular) EFMs in these networks, I introduced the concept of the atomic elementary flux mode (AEFM)—my second major computational contribution of my thesis—and provided intuition on why these pathways could overcome limitations associated with molecular EFMs. I leveraged existing work in the atom mapping field to adapt my CHMC method to enumerate AEFMs and compute their weights. (One could think of the CHMC analysis of the sphingolipid kinetic model in Chapter 2 as an atomic cycle-history Markov chain

(ACHMC) describing any of the backbone carbons shared across all lipid classes). Using this approach, I comprehensively analyzed five genome-scale metabolic models (GEMs) of increasing size and complexity. I demonstrated that AEFM enumeration was possible on networks that are computationally infeasible for molecular EFM enumeration programs after two decades of algorithmic advancements.[57,61] This breakthrough led to a structural analysis of these AEFMs to examine the types of steady state atomic pathways involved in these networks. These analyses culminated in a final carbon AEFM weight analysis of a publicly available human liver cancer cell line (HepG2) metabolic flux network, which predicted well-known and novel pathways of glutamine remodelling.

8.2 Cautionary remarks and limitations

8.2.1 Validating (atomic) elementary flux mode ((A)EFM) weights assigned under the (A)CHMC model

A question that arises from Chapters 3 and 6 is on validating (A)EFM weights assigned by the (A)CHMC model. This question can be addressed experimentally by (i) tracing single molecules (ii) in real time, (iii) across multiple reactions in a (iv) single living cell. However, it is not possible to achieve all four demands with the technology available today. While much effort has been made in this domain (e.g.[274–279]), only a subset of these methods can be used on live cells, and achieving real-time monitoring at the single molecule resolution is generally limited to single reactions. If one cannot confirm the assigned (A)EFM weights, should one be worried about basing biological predictions on the (A)CHMC model?

I argue that the Markov constraint of the (A)CHMC is a natural approach to model particle flux, and has been successsfully used to model other types of biophysical processes.[181,280,281] In contrast, mathematical-optimization-based approaches rely on no-

tions of parsimony under the assumption that individual molecules have agency to travel through a designated series of reactions forming an EFM. This assumption is biophysically implausible since molecules have no memory of the reactions preceding or following them. Until EFM and AEFM weights can be experimentally verified, I believe a Markovian approach to assigning weights is the most realistic and state-of-the-art model for modelling the steady state flows of molecules and atoms within a flux network.

8.2.2 Effects of incorrect atom mapping predictions

The work conducted in Chapter 7 was made possible by recent developments in atom mapping algorithms. These methods were critical given the hundreds of unique reactions present across the five metabolic networks. While RXNMapper boasts a 99.4% accuracy across 49,000 strongly unbalanced reactions,[177] I observed the program made several mistakes across four of the five datasets. These errors were caught based on my personal knowledge of metabolic reactions and involved ATP-independent aminotransferase reactions (Figure D.5). This problem has been observed in previous benchmarking studies of atom mapping programs.[273] Hence, I am surprised they were not part of the RXNMapper test dataset or commented on by the authors in their paper.

These atom mapping errors required rerunning my ACHMC pipeline and regenerating my analyses in Chapter 7. Comparing results from the incorrect and correct mappings provided valuable insight into problems that may occur when ACHMCs are erroneously constructed. For the HepG2 dataset, specifically, I found incorrect aminotransferase reactions caused glutamine-derived carbons to remodel into metabolites that are biologically impossible. This atom "leakage" also resulted in an explosion of ACHMC states with increased computational run times to enumerate the incorrect AEFMs and assign incorrect weights (data not shown). Moving forward, I recommend

that metabolic models link atom mapping data from curated databases such as KEGG[75] and MetaCyc[282] in their SBML files.

8.2.3 Facilitating AEFM analyses with better atom mapping data

Two major bottlenecks in running the ACHMC pipeline are compiling the list of metabolite simplified molecular input entry system (SMILES) strings and atom mappings across all metabolites and reactions in the input metabolic model. While RXN-Mapper was used to address the latter problem, I demonstrated that it could still produce incorrect atom mapping predictions. Advancements in atom mapping algorithms may reduce the number of incorrect mappings (e.g.[283]), but one must still manually compile the list of metabolite SMILES strings to input into these atom mapping prediction programs.

Rather than addressing the atom mapping problem via computational inference methods, a more practical solution may be to simply improve database search tools to extract existing atom mappings (and metabolite SMILES strings) published in literature. First, the vast majority of biochemical reactions in metabolic databases and published metabolic models already contain curated atom mappings (e.g. KEGG[75]). Second, many biochemical reactions involve small molecules with simple structures, leading to intuitive atom mappings that can be manually curated. Thus, the need for computational tools to predict these atom mappings should diminish over time as more biochemical reactions are atom mapped by expert curators. To leverage this information, I recommend that metabolite and reaction identifiers be federated across existing metabolic databases. This would facilitate the extraction of metabolite and reaction metadata because different databases often use different names to refer to the same metabolite or reaction. Depending on the database, for example, the metabolite 2-oxoglutarate is also referred to as alpha-ketogluratate or α-ketoglutarate. By unifying

these namespaces, one could envision automated database searching methods to extract curated metabolite structures and atom mappings regardless of the input metabolite or reaction name. This would significantly decrease the time needed to set up the ACHMC pipeline for publicly available or even custom metabolic models.

8.2.4 Conducting AEFM analyses today

The proposal of EFMs in 1994 sparked a flurry of research to explore the potential of this analytical framework. Aside from successful applications engineering microbial and yeast strains,[41,210,284,285] few biological advances[38,286] were made due to the EFM enumeration and flux decomposition problem.

I am hopeful that AEFMs and my ACHMC method will stimulate renewed interest in the steady state analysis of metabolic flux networks. However, I recognize there are many barriers for conducting these analyses today. The greatest challenge is the paucity of experimentally determined metabolic flux datasets. Stable isotope tracing analysis (SITA) is the premiere technique to measure fluxes, yet requires expensive equipment (mass spectrometers) reagents (stable isotope standards) and skilled operators capable of designing, executing, and interpreting these experiments. Further advancements in metabolic flux analysis are required to infer metabolic fluxes from the resulting isotopic distributions of internal metabolites. Setting this problem aside, current SITA protocols today are incapable of measuring hundreds of fluxes in large-scale networks. EFM and AEFM analyses may be applied to small-scale flux networks, yet these are typically simple enough to be analyzed by hand without the need for specialized (A)EFM enumeration and flux decomposition tools (e.g. Schwartz et al.[55]).

The high costs associated with SITA have limited ability to measure fluxes has led to a rise in computational methods to infer fluxes from other high-throughput data sources. Many of these techniques are based on flux balance analysis (FBA) with the incorpora-

tion of transcriptomics,[105–110,112,113,287] proteomics,[114–116,136,288–290] or metabolomics[288] data. The benefit of this approach is that abundance-based data is much cheaper and easier to acquire (e.g.[291–293]), with many publicly available datasets across biological conditions. One could infer a distribution of steady state fluxes from each transcript/metabolite/protein abundance dataset and perform a subsequent AEFM analysis. This approach comes with two major caveats. First, the use of abundance-based data to predict fluxes restricts them from being integrated into a multi-omic model of metabolism due to the double-use of abundance data. I also caution against estimating fluxes from omics-guided FBA since correlations between metabolic fluxes and transcript, metabolite, and protein levels is generally quite poor[294–306] leading to inaccurate flux estimates.[207]

8.3 Avenues of future work

8.3.1 Improving ACHMC scaling

Conducting AEFM analyses on GEM will require further optimizations and algorithmic improvements to my ACHMC implementation. Analysis of my computational run times in Chapter 7 revealed my implementation to scale quadratically as a function of AEFMs. After implementing the original CHMC and ACHMC methods, I acknowledge there are many improvements I could have made in hindsight to make my implementation run even faster. Here, I remark on several changes to boost AEFM extraction and flux decomposition.

The process of enumerating AEFMs and identifying their weights can be broken down into four computationally demanding steps described in Chapters 3 and 6. These are (i) constructing the atomic transition graph and ACHMC, (ii) identifying simple cycle closures, (iii) collecting simple cycle closures by AEFMs, and (iv) computing

the stationary distribution of the ACHMC. Step (i) is computationally straightforward since constructing the ACHMC is done recursively by repeatedly searching the product metabolite/atom position from a precomputed atom mapping dictionary to populate a (sparse) ACHMC transition matrix. Since Julia does not implement tail-call optimisation, this step could potentially be sped up by storing simple paths in a stack and using pushing/popping operations to enumerate the ACHMC states. I consider step (iv) to be fully optimised since solving for the stationary distribution is handled by the Julia package LinearSolve.jl which, from my tests (see Appendix C), is the fastest and most numerically stable method to compute eigenvectors of large ACHMC matrices. Numeric instability could be reduced by compressing linear ACHMC pathways to reduce the matrix size and potentially decrease the time required to compute the eigenvector. Another interesting observation is that only the steady state probabilities of ACHMC states involved in simple cycle closures are required to compute the AEFM probabilities. Provided they are proportional to one another, I ponder whether one could develop a faster method to compute a subset of eigenvector elements for specific ACHMC states.

I have found that steps (ii) and (iii) are the most computationally demanding aspect of my ACHMC implementation. I first look through each ACHMC prefix to identify simple cycles before searching this list to group those belonging to the same AEFM. Step (iii) can be readily optimized because sorting simple cycles into AEFMs is done naïvely by concatenating a reference simple cycle with itself and testing whether consecutive metabolite/atom position indices in a candidate simple cycle match to consecutive indices in that reference. If n is the number of candidate simple cycles with the same length as the reference, and k is the length of each candidate, this approach has a worst-case run time of $\mathcal{O}(k \cdot n)$. I have since learned about lexographically minimal string rotations and how I could have normalized all of my simple cycles to start with, say, the smallest numeric index. After rotating each simple cycle, matching candidate

simple cycles to a reference would run in $\mathcal{O}(n)$. Further performance gains aggregating simple cycles may be had through alignment algorithms used in sequencing-based bioinformatics. Examples include Burrows-Wheeler Transform (BWT) to efficiently store ACHMC states, or hashing functions to aggregate simple cycles corresponding to the same AEFMs.

8.3.2 Developing new methods to analyze AEFM weights

My ACHMC method opens several avenues for analyzing steady state, metabolic flux data. As mentioned in Chapter 7, AEFMs and their weights are dependent on one another due to their shared, underlying reactions. For example, several dominant glutamine-derived carbon AEFMs corresponded to aspects of the malate-aspartate shuttle in my AEFM weight analysis of the HepG2 dataset. One question is whether AEFMs correspond with traditionally defined metabolic subsystems. This problem is more complex than one would think given overlapping metabolites between subsystems and multiple comparisons problem across thousands of AEFMs and dozens of metabolic subsystems. In a naïve over-representation analysis using a hypergeometric test (data not shown), I found largely insignificant P-values between metabolites within AEFMs and KEGG subsystems, regardless of the AEFM weight.

Mapping AEFMs to metabolic subsystems leads to another problem of analyzing the joint distribution of AEFM weights within and across source metabolites. These analyses would need to consider the underlying reactions shared between AEFMs and their steady state fluxes. Coupled pathway fluxes may point to higher-order metabolic regulation of particular substrates. Combined with the dominant pathway flux hypothesis formed in Chapter 5, certain pathway fluxes may tend to dominate across biological conditions or organisms. The invariance of pathway fluxes that are constuititively active across organisms could point to the evolutionary significance or strong regulatory

forces maintaining flux through these pathways.

Yet another interesting idea is to quantify changes in (A)EFM weights as a function of changes to enzyme kinetics or other network perturbations. This is essentially a sensitivity analysis of (A)EFM weights similar in principle to metabolic control analysis (MCA). One could imagine perturbing one or more single kinetic parameters in a network, computing the steady state fluxes, and decomposing those onto (A)EFMs. Two broad questions that come to mind from this type of experiment are whether control exerted by one or more enzymes is identical across (A)EFMs , and whether enzyme kinetic parameters are "sloppy,"[133] leading to relatively stable (A)EFM weights. For the former question, one could perturb a single kinetic parameter and observe how the weights change across all (A)EFMs involving that reaction, in addition to their ranks across all network (A)EFMs. For the latter question, I would hypothesize that global kinetic perturbations to the network would result in discrete metabolic phenotypes characterized by specific (A)EFM rank weight distributions. This invariance could also be tested by running this experiment across networks parameterized according to different enzyme kinetic parameter prediction tools.[135,307,308] A discovery like this would point to the simplicity of pathway flux distributions arising from complex, kinetic models of metabolism.

I finally highlight whether dominant AEFMs can identify novel metabolic pathways or interesting variations of existing ones. These questions are highly relevant in the context of substrate utilization to characterize and engineer microbes and yeast. It is often easier to generate or infer fluxes in these organisms due to their simpler metabolic networks and predictable metabolic phenotypes, and they are often studied to optimize the production of desirable compounds. Carbon fixation pathways,[309] for example, explain the metabolism of input one carbon (C1) compounds into organic compounds. Variations of these pathways continue to be discovered such as the Gnd-Entner-Doudoroff (GED) cycle,[310] in addition to completely new metabolic processes such as the reductive

glycine pathway by Sánchez-Andrea et al. in 2020.[311] These pathways share similarities with AEFMs given C1 compounds correspond to single carbon metabolites (although some extend C1 to compounds with no carbon-carbon bonds.[312,313]) AEFM analysis of C1 metabolites may prove useful in quantifying pathway flux in organisms capable of undergoing multiple carbon fixation pathways or, perhaps, discovering new metabolic pathways altogether.

References

1. King, ZA, Lu, J, Dräger, A, Miller, P, Federowicz, S, Lerman, JA, Ebrahim, A, Palsson, BØ, Lewis, NE. BiGG Models: a platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Res. 2016;44:D515–D522. doi:10.1093/nar/gkv1049.

2. Norsigian, CJ, Pusarla, N, McConn, JL, Yurkovich, JT, Dräger, A, Palsson, BØ, King, Z. BiGG Models 2020: multi-strain genome-scale models and expansion across the phylogenetic tree. Nucleic Acids Research. 2019;48:D402–D406. doi:10.1093/nar/gkz1054.

3. Buchner, E. Alkoholische Gährung ohne Hefezellen. Berichte der deutschen chemischen Gesellschaft. 1897;30:117–124. doi:10.1002/cber.18970300121.

4. Krebs, HA, Johnson, WA. Acetopyruvic acid (alphagamma-diketovaleric acid) as an intermediate metabolite in animal tissues. Biochem J. 1937;31:772–779. doi:10.1042/bj0310772.

5. Krebs, HA, Johnson, WA. Metabolism of ketonic acids in animal tissues. Biochem J. 1937;31:645–660. doi:10.1042/bj0310645.

6. Krebs, HA, Johnson, WA. The role of citric acid in intermediate metabolism in animal tissues. Enzymologia. 1937;4:148–156. doi:10.5555/19371405235.

7. Grüning, NM, Ralser, M. Glycolysis: How a 300yr long research journey that started with the desire to improve alcoholic beverages kept revolutionizing biochemistry. Current opinion in systems biology. 2021;28:100380. doi:10.1016/j.coisb.2021.100380.

8. Horecker, B, Smyrniotis, P, Seegmiller, J. The enzymatic conversion of 6-phosphogluconate to ribulose-5-phosphate and ribose-5-phosphate. J Biol Chem. 1951;193:383–396. doi:10.1016/S0021-9258(19)52464-4.

9. Gunsalus, IC, Horecker, BL, Wood, WA. Pathways of carbohydrate metabolism in microorganisms. Bacteriol Rev. 1955;19:79–128. doi:10.1128/br.19.2.79-128.1955.

10. Ruben, S, Kamen, MD. Long-Lived radioactive carbon: C^{14}. Phys Rev. 4 1941;59:349–354. doi:10.1103/PhysRev.59.349.

11. Quayle, JR, Fuller, RC, Benson, AA, Calvin, M. Enzymatic carboxylation of ribulose diphosphate 1. Journal of the American Chemical Society. 1954;76:3610–3611. doi:10.1021/ja01642a089.

12. Bassham, JA, Benson, AA, Kay, LD, Harris, AZ, Wilson, AT, Calvin, M. The path of carbon in photosynthesis. XXI. The cyclic regeneration of carbon dioxide acceptor 1. J Am Chem Soc. 1954;76:1760–1770. doi:10.1021/ja01636a012.

13. Lehninger, AL. Phosphorylation coupled to oxidation of dihydrodiphosphopyridine nucleotide. J Biol Chem. 1951;190:345–359. doi:10.1016/S0021-9258(18)56077-4.

14. Borst, P. The aerobic oxidation of reduced diphosphopyridine nucleotide formed by glycolysis in ehrlich ascites-tumour cells. Biochim Biophys Acta. 1962;57:270–282. doi:10.1016/0006-3002(62)91120-4.

15. Dawson, A. Oxidation of cytosolic NADH formed during aerobic metabolism in mammalian cells. Trends Biochem Sci. 1979;4:171–176. doi:10.1016/0968-0004(79)90417-1.

16. Nicholson, DE. The evolution of the IUBMB-Nicholson maps. IUBMB Life. 2000;50:341–344. doi:10.1080/152165400300089295.

17. Ishii, N, Nakahigashi, K, Baba, T, Robert, M, Soga, T, Kanai, A, Hirasawa, T, Naba, M, Hirai, K, Hoque, A, et al. Multiple high-throughput analyses monitor the response of E. coli to perturbations. Science. 2007;316:593–597. doi:10.1126/science.1132067.

18. Sauer, U, Heinemann, M, Zamboni, N. Genetics. Getting closer to the whole picture. Science. 2007;316:550–551. doi:10.1126/science.1142502.

19. Jones, S, Robertson, G, Hirst, M, Bainbridge, M, Bilenky, M, Zhao, Y, Zeng, T, Euskirchen, G, Bernier, B, Varhol, R, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods. 2007;4:651–657. doi:10.1038/nmeth1068.

20. Lieberman-Aiden, E, van Berkum, NL, Williams, L, Imakaev, M, Ragoczy, T, Telling, A, Amit, I, Lajoie, BR, Sabo, PJ, Dorschner, MO, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326:289–293. doi:10.1126/science.1178746.

21. Roux, KJ, Kim, DI, Raida, M, Burke, B. A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. J Cell Biol. 2012;196:801–810. doi:10.1083/jcb.201112098.

22. Hicks, KG, Cluntun, AA, Schubert, HL, Hackett, SR, Berg, JA, Leonard, PG, Ajalla Aleixo, MA, Zhou, Y, Bott, AJ, Salvatore, SR, et al. Protein-metabolite interactomics of carbohydrate metabolism reveal regulation of lactate dehydrogenase. Science. 2023;379:996–1003. doi:10.1126/science.abm3452.

23. Westerhoff, HV, Palsson, BØ. The evolution of molecular biology into systems biology. Nat Biotechnol. 2004;22:1249–1252. doi:10.1038/nbt1020.

24. Kitano, H. Computational systems biology. Nature. 2002;420:206–210. doi:10.1038/nature01254.

25. Stelling, J, Sauer, U, Szallasi, Z, Doyle, FJ, Doyle, J. Robustness of cellular functions. Cell. 2004;118:675–685. doi:10.1016/j.cell.2004.09.008.

26. Heinrich, R, Schuster, S. *The regulation of cellular systems* (Chapman & Hall, New York, NY, 1996).

27. Kacser, H, Burns, JA. The control of flux. Symp Soc Exp Biol. 1973;27:65–104.

28. Kacser, H, Burns, JA, Fell, DA. The control of flux. Symp Soc Exp Biol. 1995;23:341–391. doi:10.1042/bst0230341.

29. Heinrich, R, Rapoport, TA. A linear steady-state treatment of enzymatic chains. Eur J Biochem. 1974;42:97–105. doi:10.1111/j.1432-1033.1974.tb03319.x.

30. Heinrich, R, Rapoport, TA. A linear steady-state treatment of enzymatic chains. Critique of the crossover theorem and a general procedure to identify interaction sites with an effector. Eur J Biochem. 1974;42:97–105. doi:10.1111/j.1432-1033.1974.tb03319.x.

31. Fell, DA, Sauro, HM. Metabolic control and its analysis: additional relationships between elasticities and control coefficients. Eur J Biochem. 1985;148:555–561. doi:10.1111/j.1432-1033.1985.tb08876.x.

32. Hadidian, Z, Hoagland, H. Chemical pacemakers: III. Activation energies of some rate-limiting components of respiratory systems. J Gen Physiol. 1941;24:339–352. doi:10.1085/jgp.24.3.339.

33. Krebs, H. Rate control of the tricarboxylic acid cycle. Advances in enzyme regulation. 1970;8:335–353. doi:10.1016/0065-2571(70)90028-2.

34. Harper, ME, Patti, ME. Metabolic terminology: what's in a name? Nat Metab. 2020;2:476–477. doi:10.1038/s42255-020-0216-7.

35. Varma, AH, Palsson, BØ. Metabolic flux balancing: basic concepts, scientific and practical use. Bio/Technology. 1994;12:994–998. doi:10.1038/nbt1094-994.

36. Orth, JD, Fleming, RMT, Palsson, BØ. Reconstruction and use of microbial metabolic networks: the core Escherichia coli metabolic model as an educational guide. EcoSal Plus. 2010;4:1128. doi:10.1128/ecosalplus.10.2.1.

37. Schuster, S, Hilgetag, C. On elementary flux modes in biochemical reaction systems at steady state. J Biol Syst. 1994;2:165–182. doi:10.1142/S0218339094000131.

38. Stelling, J, Klamt, S, Bettenbrock, K, Schuster, S, Gilles, ED. Metabolic network structure determines key aspects of functionality and regulation. Nature. 2002;420:190–193. doi:10.1038/nature01166.

39. Deutscher, D, Meilijson, I, Kupiec, M, Ruppin, E. Multiple knockout analysis of genetic robustness in the yeast metabolic network. Nat Genet. 2006;38:993–998. doi:10.1038/ng1856.

40. Behre, J, Wilhelm, T, von Kamp, A, Ruppin, E, Schuster, S. Structural robustness of metabolic networks with respect to multiple knockouts. J Theor Biol. 2008;252:433–441. doi:10.1016/j.jtbi.2007.09.043.

41. Carlson, R, Srienc, F. Fundamental Escherichia coli biochemical pathways for biomass and energy production: Identification of reactions. Biotechnol Bioeng. 2004;85:1–19. doi:10.1002/bit.10812.

42. Carlson, R, Srienc, F. Fundamental Escherichia coli biochemical pathways for biomass and energy production: creation of overall flux states. Biotechnol Bioeng. 2004;86:149–162. doi:10.1002/bit.20044.

43. Unrean, P, Trinh, CT, Srienc, F. Rational design and construction of an efficient E. coli for production of diapolycopendioic acid. Metab Eng. 2010;12:112–122. doi:10.1016/j.ymben.2009.11.002.

44. Hosack, DA, Dennis Jr, G, Sherman, BT, Lane, HC, Lempicki, RA. Identifying biological themes within lists of genes with EASE. Genome Biol. 2003;4:R70. doi:10.1186/gb-2003-4-10-r70.

45. Doniger, SW, Salomonis, N, Dahlquist, KD, Vranizan, K, Lawlor, SC, Conklin, BR. MAPPFinder: using gene ontology and GenMAPP to create a global gene-expression profile from microarray data. Genome Biol. 2003;4:R7. doi:10.1186/gb-2003-4-1-r7.

46. Wieder, C, Frainay, C, Poupin, N, Rodríguez-Mier, P, Vinson, F, Cooke, J, Lai, RP, Bundy, JG, Jourdan, F, Ebbels, T, et al. Pathway analysis in metabolomics: Recommendations for the use of over-representation analysis. PLoS Comput Biol. 2021;17:e1009105. doi:10.1371/journal.pcbi.1009105.

47. Simillion, C, Liechti, R, Lischer, HEL, Ioannidis, V, Bruggmann, R. Avoiding the pitfalls of gene set enrichment analysis with SetRank. BMC Bioinformatics. 2017;18:151. doi:10.1186/s12859-017-1571-6.

48. Subramanian, A, Tamayo, P, Mootha, VK, Mukherjee, S, Ebert, BL, Gillette, MA, Paulovich, A, Pomeroy, SL, Golub, TR, Lander, ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005;102:15545–15550. doi:10.1073/pnas.0506580102.

49. Barry, WT, Nobel, AB, Wright, FA. Significance analysis of functional categories in gene expression studies: a structured permutation approach. Bioinformatics. 2005;21:1943–1949. doi:10.1093/bioinformatics/bti260.

50. Draghici, S, Khatri, P, Tarca, AL, Amin, K, Done, A, Voichita, C, Georgescu, C, Romero, R. A systems biology approach for pathway level analysis. Genome Res. 2007;17:1537–1545. doi:10.1101/gr.6202607.

51. Grassi, M, Tarantino, B. SEMgsa: topology-based pathway enrichment analysis with structural equation models. BMC Bioinformatics. 2022;23:344. doi:10.1186/s12859022048848.

52. Nookaew, I, Meechai, A, Thammarongtham, C, Laoteng, K, Ruanglek, V, Cheevadhanarak, S, Nielsen, J, Bhumiratana, S. Identification of flux regulation coefficients from elementary flux modes: A systems biology tool for analysis of metabolic networks. Biotechnol Bioeng. 2007;97:1535–1549. doi:10.1002/bit.21339.

53. Orman, MA, Berthiaume, F, Androulakis, IP, Ierapetritou, MG. Pathway analysis of liver metabolism under stressed condition. J Theor Biol. 2011;272:131–140. doi:10.1016/j.jtbi.2010.11.042.

54. Rügen, M, Bockmayr, A, Legrand, J, Cogne, G. Network reduction in metabolic pathway analysis: Elucidation of the key pathways involved in the photoautotrophic growth of the green alga Chlamydomonas reinhardtii. Metab Eng. 2012;14:458–467. doi:10.1016/j.ymben.2012.01.009.

55. Schwartz, JM, Kanehisa, M. Quantitative elementary mode analysis of metabolic pathways: the example of yeast glycolysis. BMC Bioinformatics. 2006;7:186. doi:10.1186/1471-2105-7-186.

56. Ren, L, Sun, X, Zhang, L, Zhao, Q, Huang, H. Identification of active pathways of Chlorella protothecoides by elementary mode analysis integrated with fluxomic data. Algal Res. 2020;45:101767. doi:10.1016/j.algal.2019.101767.

57. Pfeiffer, T, Sánchez-Valdenebro, I, Nuño, JC, Montero, F, Schuster, S. METATOOL: for studying metabolic networks. Bioinformatics. 1999;15:251–257. doi:10.1093/bioinformatics/15.3.251.

58. Gagneur, J, Klamt, S. Computation of elementary modes: a unifying framework and the new binary approach. BMC Bioinformatics. 2004;5:175. doi:10.1186/1471-2105-5-175.

59. Terzer, M, Stelling, J. Large-scale computation of elementary flux modes with bit pattern trees. Bioinformatics. 2008;24:2229–2235. doi:10.1093/bioinformatics/btn401.

60. Van Klinken, JB, Willems van Dijk, K. FluxModeCalculator: an efficient tool for large-scale flux mode computation. Bioinformatics. 2016;32:1265–1266. doi:10.1093/bioinformatics/btv742.

61. Buchner, BA, Zanghellini, J. EFMlrs: a Python package for elementary flux mode enumeration via lexicographic reverse search. BMC Bioinformatics. 2021;22:547. doi:10.1186/s12859-021-04417-9.

62. Ullah, E, Yosafshahi, M, Hassoun, S. Towards scaling elementary flux mode computation. Brief Bioinform. 2020;21:1875–1885. doi:10.1093/bib/bbz094.

63. Breuer, M, Earnest, TM, Merryman, C, Wise, KS, Sun, L, Lynott, MR, Hutchison, CA, Smith, HO, Lapek, JD, Gonzalez, DJ, et al. Essential metabolism for a minimal cell. eLife. 2019;8. doi:10.7554/eLife.36842.

64. Acuña, V, Chierichetti, F, Lacroix, V, Marchetti-Spaccamela, A, Sagot, MF, Stougie, L. Modes and cuts in metabolic networks: Complexity and algorithms. Biosystems. 2009;95:51–60. doi:10.1016/j.biosystems.2008.06.015.

65. Acuña, V, Marchetti-Spaccamela, A, Sagot, MF, Stougie, L. A note on the complexity of finding and enumerating elementary modes. BioSystems. 2010;99:210–214. doi:10.1016/j.biosystems.2009.11.004.

66. Becker, SA, Palsson, BØ. Genome-scale reconstruction of the metabolic network in Staphylococcus aureus N315: an initial draft to the two-dimensional annotation. BMC Microbiol. 2005;5:8–8. doi:10.1186/1471-2180-5-8.

67. Joshi, A, Palsson, BØ. Metabolic dynamics in the human red cell. Part I–a comprehensive kinetic model. J Theor Biol. 1989;141:515–528. doi:10.1016/S0022-5193(89)80233-4.

68. Reed, JL, Vo, TD, Schilling, CH, Palsson, BØ. An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). Genome Biol. 2003;4:R54. doi:10.1186/gb-2003-4-9-r54.

69. Wang, H, Marcišauskas, S, Sánchez, BJ, Domenzain, I, Hermansson, D, Agren, R, Nielsen, J, Kerkhoven, EJ, Ouzounis, CA. RAVEN 2.0: a versatile toolbox for metabolic network reconstruction and a case study on Streptomyces coelicolor. PLoS Comput Biol. 2018;14:e1006541. doi:10.1371/journal.pcbi.1006541.

70. Heirendt, L, Arreckx, S, Pfau, T, Mendoza, SN, Richelle, A, Heinken, A, Haraldsdóttir, HS, Wachowiak, J, Keating, SM, Vlasov, V, et al. Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. Nat Protoc. 2019;14:639–702. doi:10.1038/s41596-018-0098-2.

71. Arkin, AP, Cottingham, RW, Henry, CS, Harris, NL, Stevens, RL, Maslov, S, Dehal, P, Ware, D, Perez, F, Canon, S, et al. KBase: The United States Department of Energy systems biology knowledgebase. Nat Biotechnol. 2018;36:566–569. doi:10.1038/nbt.4163.

72. Tarzi, C, Zampieri, G, Sullivan, N, Angione, C. Emerging methods for genome-scale metabolic modeling of microbial communities. Trends Endocrinol Metab. 2024;35:533–548. doi:10.1016/j.tem.2024.02.018.

73. Sayers, EW, Bolton, EE, Brister, JR, Canese, K, Chan, J, Comeau, DC, Connor, R, Funk, K, Kelly, C, Kim, S, et al. Database resources of the national center for biotechnology information. Nucleic Acids Res. 2022;50:D20–D26. doi:10.1093/nar/gkab1112.

74. Madeira, F, Madhusoodanan, N, Lee, J, Eusebi, A, Niewielska, A, Tivey, ARN, Lopez, R, Butcher, S. The EMBL-EBI job dispatcher sequence analysis tools framework in 2024. Nucleic Acids Res. 2024;52:W521–W525. doi:10.1093/nar/gkae241.

75. Kanehisa, M, Furumichi, M, Sato, Y, Kawashima, M, Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res. 2023;51:D587–D592. doi:10.1093/nar/gkac963.

76. Seaver, SMD, Liu, F, Zhang, Q, Jeffryes, J, Faria, JP, Edirisinghe, JN, Mundy, M, Chia, N, Noor, E, Beber, ME, et al. The ModelSEED biochemistry database for the integration of metabolic annotations and the reconstruction, comparison and analysis of metabolic models for plants, fungi and microbes. Nucleic Acids Res. 2021;49:D1555. doi:10.1093/nar/gkaa746.

77. Saier, MH, Reddy, VS, Moreno-Hagelsieb, G, Hendargo, KJ, Zhang, Y, Iddamsetty, V, Lam, KJK, Tian, N, Russum, S, Wang, J, et al. The transporter classification database (TCDB): 2021 update. Nucleic Acids Res. 2021;49:D461–D467. doi:10.1093/nar/gkaa1004.

78. Bernstein, DB, Sulheim, S, Almaas, E, Segrè, D. Addressing uncertainty in genome-scale metabolic model reconstruction and analysis. Genome Biol. 2021;22:64. doi:10.1186/s13059-021-02289-z.

79. Nilsson, A, Haanstra, JR, Engqvist, M, Gerding, A, Bakker, BM, Klingmüller, U, Teusink, B, Nielsen, J. Quantitative analysis of amino acid metabolism in liver cancer links glutamate excretion to nucleotide synthesis. Proc Natl Acad Sci USA. 2020;117:10294–10304. doi:10.1073/pnas.1919250117.

80. Chitpin, JG. Bioinformatics for quantitative, targeted lipidomics discovery [dissertation]. Ottawa (Canada): University of Ottawa; 2020. doi:10.20381/ruor-25020.

81. L'Annunziata, MF. *Radioactivity: introduction and history, from the quantum to quarks* (Elsevier Science, Atlanta, GA, 2016).

82. Godwin, H. Half-life of radiocarbon. Nature. 1962;195:984. doi:10.1038/195984a0.

83. Ap Rees, T, Beevers, H. Pathways of glucose dissimilation in carrot slices. Plant Physiol. 1960;35:830–838.

84. Doniach, I, Pelc, SR. Autoradiograph technique. Brit J Radiol. 1950;23:184–192. doi:10.1259/0007-1285-23-267-184.

85. Gauthier-Coles, G, Vennitti, J, Zhang, Z, Comb, WC, Xing, S, Javed, K, Bröer, A, Bröer, S. Quantitative modelling of amino acid transport and homeostasis in mammalian cells. Nat Commun. 2021;12:5282. doi:10.1038/s41467-021-25563-x.

86. Muller, RA. Radioisotope dating with a cyclotron. Science. 1977;196:489–494.

87. Antoniewicz, MR. A guide to 13C metabolic flux analysis for the cancer biologist. Exp Mol Med. 2018;50:1–13. doi:10.1038/s12276-018-0060-y.

88. Wang, L, Xing, X, Zeng, X, Jackson, SR, TeSlaa, T, Al-Dalahmah, O, Samarah, LZ, Goodwin, K, Yang, L, McReynolds, MR, et al. Spatially resolved isotope tracing reveals tissue metabolic activity. Nat Methods. 2022;19:223–230. doi:10.1038/s41592-021-01378-y.

89. Cherkaoui, S, Durot, S, Bradley, J, Critchlow, S, Dubuis, S, Masiero, MM, Wegmann, R, Snijder, B, Othman, A, Bendtsen, C, et al. A functional analysis of 180 cancer cell lines reveals conserved intrinsic metabolic programs. Mol Syst Biol. 2022;18:e11033. doi:10.15252/msb.202211033.

90. Buescher, JM, Antoniewicz, MR, Boros, LG, Burgess, SC, Brunengraber, H, Clish, CB, DeBerardinis, RJ, Feron, O, Frezza, C, Ghesquiere, B, et al. A roadmap for interpreting 13C metabolite labeling patterns from cells. Curr Opin Biotechnol. 2015;34:189–201. doi:10.1016/j.copbio.2015.02.003.

91. Smith, BN. Natural abundance of the stable isotopes of carbon in biological systems. Bioscience. 1972;22:226–231.

92. Alves, TC, Pongratz, RL, Zhao, X, Yarborough, O, Sereda, S, Shirihai, O, Cline, GW, Mason, G, Kibbey, RG. Integrated, step-wise, mass-isotopomeric flux analysis of the TCA cycle. Cell Metab. 2015;22:936–947. doi:10.1016/j.cmet.2015.08.021.

93. Savinell, JM, Palsson, BØ. Network analysis of intermediary metabolism using linear optimization. I. Development of mathematical formalism. J Theor Biol. 1992;154:421–454. doi:10.1016/S0022-5193(05)80161-4.

94. Klamt, S, Regensburger, G, Gerstl, MP, Jungreuthmayer, C, Schuster, S, Mahadevan, R, Zanghellini, J, Müller, S, Kaleta, C. From elementary flux modes to elementary flux vectors: metabolic pathway analysis with arbitrary linear flux constraints. PLoS Comput Biol. 2017;13:e1005409. doi:10.1371/journal.pcbi.1005409.

95. Urbanczik, R, Wagner, C. Functional stoichiometric analysis of metabolic networks. Bioinformatics. 2005;21:4176–4180. doi:10.1093/bioinformatics/bti674.

96. Pál, C, Papp, B, Lercher, MJ. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nat Genet. 2005;37:1372–1375. doi:10.1038/ng1686.

97. Dragosits, M, Mattanovich, D. Adaptive laboratory evolution – principles and applications for biotechnology. Microb Cell Fact. 2013;12:64. doi:10.1186/1475-2859-12-64.

98. Choe, D, Lee, JH, Yoo, M, Hwang, S, Sung, BH, Cho, S, Palsson, B, Kim, SC, Cho, BK. Adaptive laboratory evolution of a genome-reduced Escherichia coli. Nat Commun. 2019;10:935. doi:10.1038/s41467-019-08888-6.

99. Ingelman, H, Heffernan, JK, Harris, A, Brown, SD, Shaikh, KM, Saqib, AY, Pinheiro, MJ, de Lima, LA, Martinez, KR, Gonzalez-Garcia, RA, et al. Autotrophic adaptive laboratory evolution of the acetogen Clostridium autoethanogenum delivers the gas-fermenting strain LAbrini with superior growth, products, and robustness. N Biotechnol. 2024;83:1–15. doi:10.1016/j.nbt.2024.06.002.

100. Feist, AM, Palsson, BØ. The biomass objective function. Curr Opin Microbiol. 2010;13:344–349. doi:10.1016/j.mib.2010.03.003.

101. Fell, DA, Small, JR. Fat synthesis in adipose tissue. An examination of stoichiometric constraints. Biochem J. 1986;238:781–786. doi:10.1042/bj2380781.

102. Varma, A, Palsson, BØ. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. Appl Environ Microbiol. 1994;60:3724–3731. doi:10.1128/aem.60.10.3724-3731.1994.

103. Mahadevan, R, Schilling, CH. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metab Eng. 2003;5:264–276. doi:10.1016/j.ymben.2003.09.002.

104. Lewis, NE, Hixson, KK, Conrad, TM, Lerman, JA, Charusanti, P, Polpitiya, AD, Adkins, JN, Schramm, G, Purvine, SO, Lopez-Ferrer, D, et al. Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. Mol Syst Biol. 2010;6:390. doi:10.1038/msb.2010.47.

105. Åkesson, M, Förster, J, Nielsen, J. Integration of gene expression data into genome-scale metabolic models. Metab Eng. 2004;6:285–293. doi:10.1016/j.ymben.2003.12.002.

106. Becker, SA, Palsson, BØ. Context-specific metabolic networks are consistent with experiments. PLoS Comput Biol. 2008;4:e1000082–e1000082. doi:10.1371/journal.pcbi.1000082.

107. Shlomi, T, Ruppin, E, Cabili, MN, Herrgård, MJ, Palsson, BØ. Network-based prediction of human tissue-specific metabolism. Nat Biotechnol. 2008;26:1003–1010. doi:10.1038/nbt.1487.

108. Jensen, PA, Papin, JA. Functional integration of a metabolic network model and expression data without arbitrary thresholding. Bioinformatics. 2011;27:541–547. doi:10.1093/bioinformatics/btq702.

109. Colijn, C, Brandes, A, Zucker, J, Lun, DS, Weiner, B, Farhat, MR, Cheng, TY, Moody, DB, Murray, M, Galagan, JE. Interpreting expression data with metabolic flux models: predicting Mycobacterium tuberculosis mycolic acid production. PLoS Comput Biol. 2009;5:e1000489–e1000489. doi:10.1371/journal.pcbi.1000489.

110. Kim, MK, Lane, A, Kelley, JJ, Lun, DS. E-Flux2 and SPOT: validated methods for inferring intracellular metabolic flux distributions from transcriptomic data. PloS One. 2016;11:e0157101–e0157101. doi:10.1371/journal.pone.0157101.

111. Huang, Y, Mohanty, V, Dede, M, Tsai, K, Daher, M, Li, L, Rezvani, K, Chen, K. Characterizing cancer metabolism from bulk and single-cell RNA-seq data using METAFlux. Nat Commun. 2023;14:4883. doi:10.1038/s41467-023-40457-w.

112. Lee, D, Smallbone, K, Dunn, WB, Murabito, E, Winder, CL, Kell, DB, Mendes, P, Swainston, N. Improving metabolic flux predictions using absolute gene expression data. BMC Syst Biol. 2012;6:73–73. doi:10.1186/1752-0509-6-73.

113. González-Arrué, N, Inostroza, I, Conejeros, R, Rivas-Astroza, M. Phenotype-specific estimation of metabolic fluxes using gene expression data. iScience. 2023;26:106201. doi:10.1016/j.isci.2023.106201.

114. Tian, M, Reed, JL, Wren, J. Integrating proteomic or transcriptomic data into metabolic models using linear bound flux balance analysis. Bioinformatics. 2018;34:3882–3888. doi:10.1093/bioinformatics/bty445.

115. Mori, M, Hwa, T, Martin, OC, De Martino, A, Marinari, E. Constrained allocation flux balance analysis. PLoS Comput Biol. 2016;12:e1004913–e1004913. doi:10.1371/journal.pcbi.1004913.

116. Labhsetwar, P, Melo, MCR, Cole, JA, Luthey-Schulten, Z. Population FBA predicts metabolic phenotypes in yeast. PLoS Comput Biol. 2017;13:e1005728–e1005728. doi:10.1371/journal.pcbi.1005728.

117. Ferner, RE, Aronson, JK. Cato Guldberg and Peter Waage, the history of the law of mass action, and its relevance to clinical pharmacology. Br J Clin Pharmacol. 2016;81:52–55. doi:10.1111/bcp.12721.

118. Wronowska, W, Charzyńska, A, Nienałtowski, K, Gambin, A. Computational modeling of sphingolipid metabolism. BMC Syst Biol. 2015;9:47. doi:10.1186/s12918-015-0176-9.

119. Jamshidi, N, Palsson, BØ. Mass action stoichiometric simulation models: incorporating kinetics and regulation into stoichiometric models. Biophys J. 2010;98:175–185. doi:10.1016/j.bpj.2009.09.064.

120. Monti, JLE, Montes, MR, Rossi, RC. Steady-state analysis of enzymes with non-Michaelis-Menten kinetics: the transport mechanism of Na+/K+-ATPase. J Biol Chem. 2017;293:1373–1385. doi:10.1074/jbc.M117.799536.

121. Johnson, KA, Goody, RS. The Original Michaelis Constant: Translation of the 1913 Michaelis–Menten Paper. Biochemistry. 2011;50:8264–8269. doi:10.1021/bi201284u.

122. Slyke, DDV, Cullen, GE. The mode of action of urease and of enzymes in general. J Biol Chem. 1914;19:141–180. doi:10.1016/S0021-9258(18)88300-4.

123. Briggs, GE, Haldane, JBS. A note on the kinetics of enzyme action. Biochem J. 1925;19:338–339. doi:10.1042/bj0190338.

124. Lehninger, AL. *Principles of biochemistry* (Worth Publishers, Inc, New York, NY, 1982).

125. Nikerel, I, van Winden, W, van Gulik, W, Heijnen, J. Linear-logarithmic kinetics - a framework for modeling kinetics of metabolic reaction networks. Simulation News Europe. 2007;17:19–26.

126. Chang, A, Jeske, L, Ulbrich, S, Hofmann, J, Koblitz, J, Schomburg, I, Neumann-Schaal, M, Jahn, D, Schomburg, D. BRENDA, the ELIXIR core data resource in 2021: new developments and updates. Nucleic Acids Res. 2021;49:D498–D508. doi:10.1093/nar/gkaa1025.

127. Wittig, U, Kania, R, Golebiewski, M, Rey, M, Shi, L, Jong, L, Algaa, E, Weidemann, A, Sauer-Danzwith, H, Mir, S, et al. SABIO-RK-database for biochemical reaction kinetics. Nucleic Acids Res. 2012;40:D790–D796. doi:10.1093/nar/gkr1046.

128. Saa, P, Nielsen, LK. A general framework for thermodynamically consistent parameterization and efficient sampling of enzymatic reactions. PLoS Comput Biol. 2015;11:e1004195. doi:10.1371/journal.pcbi.1004195.

129. Saa, PA, Nielsen, LK. Construction of feasible and accurate kinetic models of metabolism: A Bayesian approach. Sci Rep. 2016;6:29635–29635. doi:10.1038/srep29635.

130. Khodayari, A, Maranas, CD. A genome-scale Escherichia coli kinetic metabolic model k-ecoli457 satisfying flux data for multiple mutant strains. Nat Commun. 2016;7:13806. doi:10.1038/ncomms13806.

131. Fröhlich, F, Kaltenbacher, B, Theis, FJ, Hasenauer, J, Stelling, J. Scalable parameter estimation for genome-scale biochemical reaction networks. PLoS Comp Biol. 2017;13:e1005331. doi:10.1371/journal.pcbi.1005331.

132. Choudhury, S, Moret, M, Salvy, P, Weilandt, D, Hatzimanikatis, V, Miskovic, L. Reconstructing kinetic models for dynamical studies of metabolism using generative adversarial networks. Nat Mach. 2022;4:710–719. doi:10.1101/2022.01.06.475020.

133. Gutenkunst, RN, Waterfall, JJ, Casey, FP, Brown, KS, Myers, CR, Sethna, JP, Arkin, AP. Universally sloppy parameter sensitivities in systems biology models. PLoS Comput Biol. 2007;3:1871–1878. doi:10.1371/journal.pcbi.0030189.

134. Srinivasan, K, Friedman, BA, Etxeberria, A, Huntley, MA, van der Brug, MP, Foreman, O, Paw, JS, Modrusan, Z, Beach, TG, Serrano, GE, et al. Alzheimer's patient microglia exhibit enhanced aging and unique transcriptional activation. Cell Rep. 2020;31:107843. doi:10.1016/j.celrep.2020.107843.

135. Heckmann, D, Lloyd, CJ, Mih, N, Ha, Y, Zielinski, DC, Haiman, ZB, Desouki, AA, Lercher, MJ, Palsson, BØ. Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. Nat Commun. 2018;9:5252. doi:10.1038/s41467-018-07652-6.

136. Arend, M, Zimmer, D, Xu, R, Sommer, F, Mühlhaus, T, Nikoloski, Z. Proteomics and constraint-based modelling reveal enzyme kinetic properties of Chlamydomonas reinhardtii on a genome scale. Nat Commun. 2023;14:4781–4781. doi:10.1038/s41467-023-40498-1.

137. Heinonen, M, Osmala, M, Mannerström, H, Wallenius, J, Kaski, S, Rousu, J, Lähdesmäki, H. Bayesian metabolic flux analysis reveals intracellular flux couplings. Bioinformatics. 2019;35:i548–i557. doi:10.1093/bioinformatics/btz315.

138. Backman, TWH, Schenk, C, Radivojevic, T, Ando, D, Singh, J, Czajka, JJ, Costello, Z, Keasling, JD, Tang, Y, Akhmatskaya, E, et al. BayFlux: a Bayesian method to quantify metabolic fluxes and their uncertainty at the genome scale. PLoS Comput Biol. 2023;19. doi:10.1371/journal.pcbi.1011111.

139. Westerhoff, HV, Groen, AK, Wanders, RJ. Modern theories of metabolic control and their applications. Biosci Rep. 1984;4:1–22. doi:10.1007/BF01120819.

140. Kacser, H, Burns, JA. Molecular democracy: who shares the controls? Biochem Soc Trans. 1979;7:1149–1160. doi:10.1042/bst0071149.

141. Reder, C. Metabolic control theory: a structural approach. J Theor Biol. 1988;135:175–201. doi:10.1016/S0022-5193(88)80073-0.

142. Westerhoff, HV, Chen, YD. How do enzyme activities control metabolite concentrations? Eur J Biochem. 1984;142:425–430. doi:10.1111/j.1432-1033.1984.tb08304.x.

143. Milner, PC. The possible mechanisms of complex reactions involving consecutive steps. J Electrochem Soc. 1964;111:228–232. doi:10.1149/1.2426089.

144. Clarke, BL. Complete set of steady states for the general stoichiometric dynamical system. The Journal of chemical physics. 1981;75:4970–4979. doi:10.1063/1.441885.

145. CLARKE, BL. Stoichiometric network analysis. Cell Biophys. 1988;12:237–253. doi:10.1007/BF02918360.

146. Zanghellini, J, Ruckerbauer, DE, Hanscho, M, Jungreuthmayer, C. Elementary flux modes in a nutshell: Properties, calculation and applications. Biotechnol J. 2013;8:1009–1016. doi:10.1002/biot.201200269.

147. Schilling, CH, Letscher, D, Palsson, BØ. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. J Theor Biol. 2000;203:229–248. doi:10.1006/jtbi.2000.1073.

148. Wortel, MT, Peters, H, Hulshof, J, Teusink, B, Bruggeman, FJ. Metabolic states with maximal specific rate carry flux through an elementary flux mode. FEBS J. 2014;281:1547–1555. doi:10.1111/febs.12722.

149. Bruggeman, FJ, Remeijer, M, Droste, M, Salinas, L, Wortel, M, Planqué, R, Sauro, HM, Teusink, B, Westerhoff, HV. Whole-cell metabolic control analysis. BioSystems. 2023;234:105067. doi:10.1016/j.biosystems.2023.105067.

150. Li, S, Huang, D, Li, Y, Wen, J, Jia, X. Rational improvement of the engineered isobutanol-producing Bacillus subtilis by elementary mode analysis. Microb Cell Fact. 2012;11:101. doi:10.1186/1475-2859-11-101.

151. Averesch, NJ, Krömer, JO. Tailoring strain construction strategies for muconic acid production in S. cerevisiae and E. coli. Metab Eng Commun. 2014;1:19–28. doi:10.1016/j.meteno.2014.09.001.

152. Poblete-Castro, I, Binger, D, Rodrigues, A, Becker, J, Martins dos Santos, VA, Wittmann, C. In-silico-driven metabolic engineering of Pseudomonas putida for enhanced production of poly-hydroxyalkanoates. Metab Eng. 2013;15:113–123. doi:10.1016/j.ymben.2012.10.004.

153. Melzer, G, Esfandabadi, ME, Franco-Lara, E, Wittmann, C. Flux Design: in silico design of cell factories based on correlation of pathway fluxes to desired properties. BMC Syst Biol. 2009;3:120. doi:10.1186/1752-0509-3-120.

154. Rezola, A, Pey, J, de Figueiredo, LF, Podhorski, A, Schuster, S, Rubio, A, Planes, FJ. Selection of human tissue-specific elementary flux modes using gene expression data. Bioinformatics. 2013;29:2009–2016. doi:10.1093/bioinformatics/btt328.

155. Papin, JA, Price, ND, Wiback, SJ, Fell, DA, Palsson, BØ. Metabolic pathways in the post-genome era. Trends Biochem Sci. 2003;28:250–258. doi:10.1016/S0968-0004(03)00064-1.

156. Chan, SHJ, Ji, P. Decomposing flux distributions into elementary flux modes in genome-scale metabolic networks. Bioinformatics. 2011;27:2256–2262. doi:10.1093/bioinformatics/btr367.

157. Carbonell, P, Fichera, D, Pandit, SB, Faulon, JL. Enumerating metabolic pathways for the production of heterologous target chemicals in chassis organisms. BMC Syst Biol. 2012;6:10. doi:10.1186/1752-0509-6-10.

158. Mores, W, Bhonsale, SS, Logist, F, Van Impe, JFM. Accelerated enumeration of extreme rays through a positive-definite elementarity test. Bioinformatics. 2024:btae723. doi:10.1093/bioinformatics/btae723.

159. Wiback, SJ, Mahadevan, R, Palsson, BØ. Reconstructing metabolic flux vectors from extreme pathways: defining the alpha-spectrum. J Theor Biol. 2003;224:313–324. doi:10.1016/s0022-5193(03)00168-1.

160. Klamt, S, Stelling, J, Ginkel, M, Gilles, ED. FluxAnalyzer: exploring structure, pathways, and flux distributions in metabolic networks on interactive flux maps. Bioinformatics. 2003;19:261–269. doi:10.1093/bioinformatics/19.2.261.

161. Klamt, S, Gilles, ED. Minimal cut sets in biochemical reaction networks. Bioinformatics. 2004;20:226–234. doi:10.1093/bioinformatics/btg395.

162. Klamt, S. Generalized concept of minimal cut sets in biochemical networks. BioSystems. 2006;83:233–247. doi:10.1016/j.biosystems.2005.04.009.

163. Gerstl, MP, Klamt, S, Jungreuthmayer, C, Zanghellini, J. Exact quantification of cellular robustness in genome-scale metabolic networks. Bioinformatics. 2016;32:730–737. doi:10.1093/bioinformatics/btv649.

164. Ballerstein, K, von Kamp, A, Klamt, S, Haus, UU. Minimal cut sets in a metabolic network are elementary modes in a dual network. Bioinformatics. 2012;28:381–387. doi:10.1093/bioinformatics/btr674.

165. Röhl, A, Riou, T, Bockmayr, A, Wren, J. Computing irreversible minimal cut sets in genome-scale metabolic networks via flux cone projection. Bioinformatics. 2019;35:2618–2625. doi:10.1093/bioinformatics/bty1027.

166. Klamt, S, Mahadevan, R, von Kamp, A. Speeding up the core algorithm for the dual calculation of minimal cut sets in large metabolic networks. BMC Bioinformatics. 2020;21:510. doi:10.1186/s12859-020-03837-3.

167. Mahout, M, Carlson, RP, Simon, L, Peres, S. Logic programming-based minimal cut sets reveal consortium-level therapeutic targets for chronic wound infections. NPJ Syst Biol Appl. 2024;10:34. doi:10.1038/s41540-024-00360-6.

168. Clement, TJ, Baalhuis, EB, Teusink, B, Bruggeman, FJ, Planqué, R, de Groot, DH. Unlocking elementary conversion modes: ecmtool unveils all capabilities of metabolic networks. Patterns. 2021;2:100177. doi:10.1016/j.patter.2020.100177.

169. Buchner, B, Clement, TJ, de Groot, DH, Zanghellini, J, Cowen, L. ecmtool: fast and memory-efficient enumeration of elementary conversion modes. Bioinformatics. 2023;39. doi:10.1093/bioinformatics/btad095.

170. Song, HS, Ramkrishna, D. Reduction of a set of elementary modes using yield analysis. Biotechnol Bioeng. 2009;102:554–568. doi:10.1002/bit.22062.

171. Luo, H, Li, P, Ji, B, Nielsen, J. Modeling the metabolic dynamics at the genome-scale by optimized yield analysis. Metabolic engineering. 2023;75:119–130. doi:10.1016/j.ymben.2022.12.001.

172. Pey, J, Theodoropoulos, C, Rezola, A, Rubio, A, Cascante, M, Planes, FJ. Do elementary flux modes combine linearly at the "atomic" level? Integrating tracer-based metabolomics data and elementary flux modes. BioSystems. 2011;105:140–146. doi:10.1016/j.biosystems.2011.04.005.

173. Pey, J, Prada, J, Beasley, JE, Planes, FJ. Path finding methods accounting for stoichiometry in metabolic networks. Genome Biol. 2011;12:R49. doi:10.1186/gb-2011-12-5-r49.

174. Pey, J, Planes, FJ, Beasley, JE. Refining carbon flux paths using atomic trace data. Bioinformatics. 2014;30:975–980. doi:10.1093/bioinformatics/btt653.

175. Pey, J, Tobalina, L, de Cisneros, JPJ, Planes, FJ. A network-based approach for predicting key enzymes explaining metabolite abundance alterations in a disease phenotype. BMC Syst Biol. 2013;7:62. doi:10.1186/1752-0509-7-62.

176. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci. 1988;28:31–36. doi:10.1021/ci00057a005.

177. Schwaller, P, Hoover, B, Reymond, JL, Strobelt, H, Laino, T. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. Sci Adv. 2021;7:eabe4166–eabe4166. doi:10.1126/sciadv.abe4166.

178. Kim, S, Chen, J, Cheng, T, Gindulyte, A, He, J, He, S, Li, Q, Shoemaker, BA, Thiessen, PA, Yu, B, et al. PubChem 2023 update. Nucleic Acids Res. 2023;51:D1373–D1380. doi:10.1093/nar/gkac956.

179. Wilkinson, DJ. *Stochastic modelling for systems biology* (Chapman & Hall, New York, NY, 2018).

180. Gillespie, DT. Exact stochastic simulation of coupled chemical reactions. J Phys Chem. 1977;81:2340–2361. doi:10.1021/j100540a008.

181. Conrad, ND, Weber, M, Schütte, C. Finding dominant structures of nonreversible Markov processes. Multiscale Model Simul. 2016;14:1319–1340. doi:10.1137/15M1032272.

182. Chitpin, JG, Perkins, TJ. A Markov constraint to uniquely identify elementary flux mode weights in unimolecular metabolic networks. J Theor Biol. 2023;575:111632. doi:10.1016/j.jtbi.2023.111632.

183. Podrini, C, Rowe, I, Pagliarini, R, Costa, ASH, Chiaravalli, M, Di Meo, I, Kim, H, Distefano, G, Tiranti, V, Qian, F, et al. Dissection of metabolic reprogramming in polycystic kidney disease reveals coordinated rewiring of bioenergetic pathways. Commun Biol. 2018;1:194. doi:10.1038/s42003-018-0200-x.

184. Chen, WW, Freinkman, E, Wang, T, Birsoy, K, Sabatini, DM. Absolute quantification of matrix metabolites reveals the dynamics of mitochondrial metabolism. Cell. 2016;166:1324–1337.e11. doi:10.1016/j.cell.2016.07.040.

185. Lau, AN, Li, Z, Danai, LV, Westermark, AM, Darnell, AM, Ferreira, R, Gocheva, V, Sivanand, S, Lien, EC, Sapp, KM, et al. Dissecting cell-type-specific metabolism in pancreatic ductal adenocarcinoma. eLife. 2020;9:e56782. doi:10.7554/eLife.56782.

186. Oftadeh, O, Salvy, P, Masid, M, Curvat, M, Miskovic, L, Hatzimanikatis, V. A genome-scale metabolic model of Saccharomyces cerevisiae that integrates expression constraints and reaction thermodynamics. Nat Commun. 2021;12:4790. doi:10.1038/s41467-021-25158-6.

187. diCenzo, GC, Tesi, M, Pfau, T, Mengoni, A, Fondi, M. Genome-scale metabolic reconstruction of the symbiosis between a leguminous plant and a nitrogen-fixing bacterium. Nat Commun. 2020;11:2574. doi:10.1038/s41467-020-16484-2.

188. Brunk, E, Sahoo, S, Zielinski, DC, Altunkaya, A, Dräger, A, Mih, N, Gatto, F, Nilsson, A, Gonzalez, GAP, Aurich, MK, et al. Recon3D: A resource enabling a three-dimensional view of gene variation in human metabolism. Nat Biotechnol. 2018;36:272–281. doi:10.1038/nbt.4072.

189. Duarte, NC, Becker, SA, Jamshidi, N, Thiele, I, Mo, ML, Vo, TD, Srivas, R, Palsson, BØ. Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci USA. 2007;104:1777–1782. doi:10.1073/pnas.0610772104.

190. Sánchez, BJ, Li, F, Kerkhoven, EJ, Nielsen, J. SLIMEr: Probing flexibility of lipid metabolism in yeast with an improved constraint-based modeling framework. BMC Syst Biol. 2019;13:4. doi:10.1186/s12918-018-0673-8.

191. De Martino, D, MC Andersson, A, Bergmiller, T, Guet, CC, Tkačik, G. Statistical mechanics for metabolic networks during steady state growth. Nat Commun. 2018;9:2988. doi:10.1038/s41467-018-05417-9.

192. Fernandez-de-Cossio-Diaz, J, Leon, K, Mulet, R. Characterizing steady states of genome-scale metabolic networks in continuous cell cultures. PLoS Comput Biol. 2017;13:e1005835. doi:10.1371/journal.pcbi.1005835.

193. Schwartz, JM, Otokuni, H, Akutsu, T, Nacher, JC. Probabilistic controllability approach to metabolic fluxes in normal and cancer tissues. Nat Commun. 2019;10:2725. doi:10.1038/s41467-019-10616-z.

194. Lee, WD, Mukha, D, Aizenshtein, E, Shlomi, T. Spatial-fluxomics provides a subcellular-compartmentalized view of reductive glutamine metabolism in cancer cells. Nat Commun. 2019;10:1351. doi:10.7554/eLife.56782.

195. Srivastava, M, Bencurova, E, Gupta, SK, Weiss, E, Löffler, J, Dandekar, T. Aspergillus fumigatus challenged by human dendritic cells: Metabolic and regulatory pathway responses testify a tight battle. Front Cell Infect Microbiol. 2019;9:168. doi:10.3389/fcimb.2019.00168.

196. Hunt, KA, Jennings, RM, Inskeep, WP, Carlson, RP. Multiscale analysis of autotroph-heterotroph interactions in a high-temperature microbial community. PLoS Comput Biol. 2018;14:e1006431. doi:10.1371/journal.pcbi.1006431.

197. Müller, S, Széliová, D, Zanghellini, J. Elementary vectors and autocatalytic sets for resource allocation in next-generation models of cellular growth. PLoS Comput Biol. 2022;18:e1009843. doi:10.1371/journal.pcbi.1009843.

198. Narisetty, V, Cox, R, Bommareddy, R, Agrawal, D, Ahmad, E, Pant, KK, Chandel, AK, Bhatia, SK, Kumar, D, Binod, P, et al. Valorisation of xylose to renewable fuels and chemicals, an essential step in augmenting the commercial viability of lignocellulosic biorefineries. Sustain Energy Fuels. 2021;6:29–65. doi:10.1039/d1se00927c.

199. Klamt, S, Stelling, J. Combinatorial complexity of pathway analysis in metabolic networks. Mol Biol Rep. 2002;29:233–236. doi:10.1023/a:1020390132244.

200. Hunt, KA, Folsom, JP, Taffs, RL, Carlson, RP. Complete enumeration of elementary flux modes through scalable demand-based subnetwork definition. Bioinformatics. 2014;30:1569–1578. doi:10.1093/bioinformatics/btu021.

201. Watson, MR. Metabolic maps for the Apple II. Biochem Soc Trans. 1984;12:1093–1094. doi:10.1042/bst0121093.

202. Schuetz, R, Kuepfer, L, Sauer, U. Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. Mol Syst Biol. 2007;3:119. doi:10.1038/msb4100162.

203. Schuster, R, Schuster, S. Refined algorithm and computer program for calculating all non–negative fluxes admissible in steady states of biochemical reaction systems with or without some flux rates fixed. Bioinformatics. 1993;9:79–85. doi:10.1093/bioinformatics/9.1.79.

204. Jenior, ML, Moutinho Jr, TJ, Dougherty, BV, Papin, JA. Transcriptome-guided parsimonious flux analysis improves predictions with metabolic networks in complex environments. PLoS Comput Biol. 2020;16:e1007099. doi:10.1371/journal.pcbi.1007099.

205. Ohno, S, Uematsu, S, Kuroda, S. Quantitative metabolic fluxes regulated by trans-omic networks. Biochem J. 2022;479:787–804. doi:10.1042/BCJ20210596.

206. Uematsu, S et al. Multi-omics-based label-free metabolic flux inference reveals obesity-associated dysregulatory mechanisms in liver glucose metabolism. iScience. 2022;25:103787. doi:10.1016/j.isci.2022.103787.

207. Machado, D, Herrgård, M. Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. PLoS Comput Biol. 2014;10:e1003580. doi:10.1371/journal.pcbi.1003580.

208. Beguerisse-Díaz, M, Bosque, G, Oyarzún, D, Picó, J, Barahona, M. Flux-dependent graphs for metabolic networks. NPJ Syst Biol Appl. 2018;4:32. doi:10.1038/s41540-018-0067-y.

209. Ruckerbauer, DE, Jungreuthmayer, C, Zanghellini, J. Predicting genetic engineering targets with elementary flux mode analysis: a review of four current methods. Nat Biotechnol. 2015;32:534–546. doi:10.1016/j.nbt.2015.03.017.

210. Schuster, S, Dandekar, T, Fell, DA. Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. Trends Biotechnol. 1999;17:53–60. doi:10.1016/s0167-7799(98)01290-6.

211. Bonarius, HP, Schmid, G, Tramper, J. Flux analysis of underdetermined metabolic networks: the quest for the missing constraints. Trends Biotechnol. 1997;15:308–314. doi:10.1016/S0167-7799(97)01067-6.

212. Shao, M, Kingsford, C. Theory and a heuristic for the minimum path flow decomposition problem. IEEE/ACM Trans Comput Biol Bioinform. 2019;16:658–670. doi:10.1109/TCBB.2017.2779509.

213. Ma, C, Kingsford, C. Deriving ranges of optimal estimated transcript expression due to non-identifiability. J Comput Biol. 2022;29:121–139. doi:10.1089/cmb.2021.0444.

214. Heinken, A, Hertel, J, Acharya, G, Ravcheev, DA, Nyga, M, Okpala, OE, Hogan, M, Magnúsdóttir, S, Martinelli, F, Nap, B, et al. Genome-scale metabolic reconstruction of 7,302 human microorganisms for personalized medicine. Nat Biotechnol. 2023;41:1320–1331. doi:10.1038/s41587-022-01628-0.

215. Von Kamp, A, Thiele, S, Hädicke, O, Klamt, S. Use of CellNetAnalyzer in biotechnology and metabolic engineering. J Biotechnol. 2017;261:221–228. doi:10.1016/j.jbiotec.2017.05.001.

216. Barabási, AL, Albert, R. Emergence of scaling in random networks. Science. 1999;286:509–512. doi:10.1126/science.286.5439.509.

217. Broido, AD, Clauset, A. Scale-free networks are rare. Nat Comms. 2019;10:1017. doi:10.1038/s41467-019-08746-5.

218. Almaas, E, Vicsek, T, Oltvai, ZN, Barabási, AL. Global organization of metabolic fluxes in the bacterium Escherichia coli. Nature. 2004;427:839–843. doi:10.1038/nature02289.

219. Pinto, CM, Mendes Lopes, A, Machado, JT. A review of power laws in real life phenomena. Commun Nonlinear Sci Numer Simul. 2012;17:3558–3578. doi:10.1016/j.cnsns.2012.01.013.

220. Bryson, MC. Heavy-tailed distributions: properties and tests. Technometrics. 1974;16:61–68. doi:10.2307/1267493.

221. Shampine, LF, Reichelt, MW. The MATLAB ODE suite. SIAM Journal on Scientific Computing. 1997;18:1–22. doi:10.1137/S1064827594276424.

222. Gillespie, CS. Fitting heavy tailed distributions: The poweRlaw package. J Stat Softw. 2015;64:1–16. doi:10.18637/jss.v064.i02.

223. Clauset, A, Shalizi, CR, Newman, MEJ. Power-law distributions in empirical data. SIAM Rev. 2009;51:661–703. doi:10.48550/arXiv.0706.1062.

224. Moivre, Ad. *The doctrine of chances, or a method of calculating the probability of events in play* (1756).

225. Park, JS, Burckhardt, CJ, Lazcano, R, Solis, LM, Isogai, T, Li, L, Chen, CS, Gao, B, Minna, JD, Bachoo, R, et al. Mechanical regulation of glycolysis via cytoskeleton architecture. Nature. 2020;578:621–626. doi:10.1038/s41586-020-1998-1.

226. Stern, A, Fokra, M, Sarvin, B, Alrahem, AA, Lee, WD, Aizenshtein, E, Sarvin, N, Shlomi, T. Inferring mitochondrial and cytosolic metabolism by coupling isotope tracing and deconvolution. Nat Commun. 2023;14:7525–7525. doi:10.1038/s41467-023-42824-z.

227. Wang, R, Yin, Y, Li, J, Wang, H, Lv, W, Gao, Y, Wang, T, Zhong, Y, Zhou, Z, Cai, Y, et al. Global stable-isotope tracing metabolomics reveals system-wide metabolic alternations in aging Drosophila. Nat Commun. 2022;13:3518. doi:10.1038/s41467-022-31268-6.

228. Zampieri, M, Hörl, M, Hotz, F, Müller, NF, Sauer, U. Regulatory mechanisms underlying coordination of amino acid and glucose catabolism in Escherichia coli. Nat Commun. 2019;10:3354. doi:10.1038/s41467-019-11331-5.

229. Kochanowski, K, Okano, H, Patsalo, V, Williamson, J, Sauer, U, Hwa, T. Global coordination of metabolic pathways in Escherichia coli by active and passive regulation. Mol Syst Biol. 2021;17:e10064. doi:10.15252/msb.202010064.

230. McKinlay, JB, Harwood, CS. Carbon dioxide fixation as a central redox cofactor recycling mechanism in bacteria. Proc Natl Acad Sci USA. 2010;107:11669–11675. doi:10.1073/pnas.1006175107.

231. Franchina, DG, Kurniawan, H, Grusdat, M, Binsfeld, C, Guerra, L, Bonetti, L, Soriano-Baguet, L, Ewen, A, Kobayashi, T, Farinelle, S, et al. Glutathione-dependent redox balance characterizes the distinct metabolic properties of follicular and marginal zone B cells. Nat Commun. 2022;13:1789. doi:10.1038/s41467-022-29426-x.

232. Basan, M, Hui, S, Okano, H, Zhang, Z, Shen, Y, Williamson, JR, Hwa, T. Overflow metabolism in Escherichia coli results from efficient proteome allocation. Nature. 2015;528:99–104. doi:10.1038/nature15765.

233. Vander Heiden, MG, Cantley, LC, Thompson, CB. Understanding the Warburg Effect: the metabolic requirements of cell proliferation. Science. 2009;324:1029–1033. doi:10.1126/science.1160809.

234. Endo, Y, Kanno, T, Nakajima, T. Fatty acid metabolism in T-cell function and differentiation. Int Immunol. 2022;34:579–587. doi:10.1093/intimm/dxac025.

235. Jackson, BT, Finley, LWS. Metabolic regulation of the hallmarks of stem cell biology. Cell Stem Cell. 2024;31:161–180. doi:10.1111/joim.12247.

236. Schuetz, R, Zamboni, N, Zampieri, M, Heinemann, M, Sauer, U. Multidimensional optimality of microbial metabolism. Science. 2012;336:601–604. doi:10.1126/science.1216882.

237. Yu, JSL, Correia-Melo, C, Zorrilla, F, Herrera-Dominguez, L, Wu, MY, Hartl, J, Campbell, K, Blasche, S, Kreidl, M, Egger, AS, et al. Microbial communities form rich extracellular metabolomes that foster metabolic interactions and promote drug tolerance. Nat Microbiol. 2022;7:542–555. doi:10.1038/s41564-022-01072-5.

238. Gustafsson, J, Roshanzamir, F, Hagnestål, A, Patel, SM, Daudu, OI, Becker, DF, Robinson, JL, Nielsen, J. Metabolic collaboration between cells in the tumor microenvironment has a negligible effect on tumor growth. Innovation. 2024;5:100583. doi:10.1016/j.xinn.2024.100583.

239. Emwas, AH, Szczepski, K, Al-Younis, I, Lachowicz, JI, Jaremko, M. Fluxomics - new metabolomics approaches to monitor metabolic pathways. Front Pharmacol. 2022;13:805782. doi:10.3389/fphar.2022.805782.

240. Sanford, K, Soucaille, P, Whited, G, Chotani, G. Genomics to fluxomics and physiomics – pathway engineering. Curr Opin Microbiol. 2002;5:318–322. doi:10.1016/s1369-5274(02)00318-1.

241. Ahn, E, Kumar, P, Mukha, D, Tzur, A, Shlomi, T. Temporal fluxomics reveals oscillations in TCA cycle flux throughout the mammalian cell cycle. Mol Syst Biol. 2017;13:953. doi:10.15252/msb.20177763.

242. Wu, C, Herold, RA, Knoshaug, EP, Wang, B, Xiong, W, Laurens, LML. Fluxomic analysis reveals central carbon metabolism adaptation for diazotroph Azotobacter vinelandii ammonium excretion. Sci Rep. 2019;9:13209. doi:10.1038/s41598-019-54633-w.

243. Kaste, JAM, Shachar-Hill, Y. Accurate flux predictions using tissue-specific gene expression in plant metabolic modeling. Bioinformatics. 2023;39:btad186. doi:10.1093/bioinformatics/btad186.

244. Lee, JY, Han, Y, Styczynski, MP. Towards inferring absolute concentrations from relative abundance in time-course GC-MS metabolomics data. Mol Omics. 2023;19:126–136. doi:10.1039/d2mo00168c.

245. Gelbach, PE, Cetin, H, Finley, SD. Flux sampling in genome-scale metabolic modeling of microbial communities. BMC Bioinformatics. 2024;25:45. doi:10.1186/s12859-024-05655-3.

246. Robinson, JL, Kocabaş, P, Wang, H, Cholley, PE, Cook, D, Nilsson, A, Anton, M, Ferreira, R, Domenzain, I, Billa, V, et al. An atlas of human metabolism. Sci Signal. 2020;13:eaaz1482. doi:10.1126/scisignal.aaz1482.

247. Wang, H, Robinson, JL, Kocabas, P, Gustafsson, J, Anton, M, Cholley, PE, Huang, S, Gobom, J, Svensson, T, Uhlen, M, et al. Genome-scale metabolic network reconstruction of model animals as a platform for translational research. Proc Natl Acad Sci USA. 2021;118:e2102344118. doi:10.1073/pnas.2102344118.

248. Nobile, MS, Coelho, V, Pescini, D, Damiani, C. Accelerated global sensitivity analysis of genome-wide constraint-based metabolic models. BMC Bioinformatics. 2021;22:78. doi:10.1186/s12859-021-04002-0.

249. Fakih, I, Got, J, Robles-Rodriguez, CE, Siegel, A, Forano, E, Muñoz-Tamayo, R. Dynamic genome-based metabolic modeling of the predominant cellulolytic rumen bacterium Fibrobacter succinogenes S85. mSystems. 2023;8:e0102722. doi:10.1128/msystems.01027-22.

250. Mpabanzi, L, Wainwright, J, Boonen, B, van Eijk, H, Dhar, D, Karssemeijer, E, Dejong, CHC, Jalan, R, Schwartz, JM, Olde Damink, SWM, et al. Fluxomics reveals cellular and molecular basis of increased renal ammoniagenesis. NPJ Syst Biol Appl. 2022;8:49. doi:10.1038/s41540-022-00257-2.

251. Toya, Y, Shimizu, H. Metabolic pathway engineering for the non-growth-associated succinate production in Escherichia coli based on flux solution space. J Biosci Bioeng. 2022;134:29–33. doi:10.1016/j.jbiosc.2022.04.008.

252. Matsuda, F, Furusawa, C, Kondo, T, Ishii, J, Shimizu, H, Kondo, A. Engineering strategy of yeast metabolism for higher alcohol production. Microb Cell Fact. 2011;10:70. doi:10.1186/1475-2859-10-70.

253. Zamboni, N, Saghatelian, A, Patti, GJ. Defining the metabolome: size, flux, and regulation. Mol Cell. 2015;58:699–706. doi:10.1016/j.molcel.2015.04.021.

254. Ye, Y, Yu, B, Wang, H, Yi, F. Glutamine metabolic reprogramming in hepatocellular carcinoma. Front Mol Biosci. 2023;10:1242059–1242059. doi:10.3389/fmolb.2023.1242059.

255. Arnold, PK, Jackson, BT, Paras, KI, Brunner, JS, Hart, ML, Newsom, OJ, Alibeckoff, SP, Endress, J, Drill, E, Sullivan, LB, et al. A non-canonical tricarboxylic acid cycle underlies cellular identity. Nature. 2022;603:477–481. doi:10.1038/s41586-022-04475-w.

256. Bordbar, A, Jamshidi, N, Palsson, BØ. iAB-RBC-283: a proteomically derived knowledge-base of erythrocyte metabolism that can be used to simulate its physiological and patho-physiological states. BMC Syst Biol. 2011;5:110–110. doi:10.1186/1752-0509-5-110.

257. Thiele, I, Vo, TD, Price, ND, Palsson, BØ. Expanded metabolic reconstruction of Helicobacter pylori (IT341 GSM/GPR): an in silico genome-scale characterization of single- and double-deletion mutants. J Bacteriol. 2005;187:5818–5830. doi:10.1128/jb.187.16.5818-5830.2005.

258. Duan, L, Cooper, DE, Scheidemantle, G, Locasale, JW, Kirsch, DG, Liu, X. 13C tracer analysis suggests extensive recycling of endogenous CO2 in vivo. Cancer Metab. 2022;10:1–11.

259. Yoo, HC, Yu, YC, Sung, Y, Han, JM. Glutamine reliance in cell metabolism. Exp Mol Med. 2020;52:1496–1516. doi:10.1038/s12276-020-00504-8.

260. Kamp, Av, Schuster, S. Metatool 5.0: fast and flexible elementary modes analysis. Bioinformatics. 2006;22:1930–1931. doi:10.1093/bioinformatics/btl267.

261. De Figueiredo, LF, Podhorski, A, Rubio, A, Kaleta, C, Beasley, JE, Schuster, S, Planes, FJ. Computing the shortest elementary flux modes in genome-scale metabolic networks. Bioinformatics. 2009;25:3158–3165. doi:10.1093/bioinformatics/btp564.

262. Gerstl, MP, Ruckerbauer, DE, Mattanovich, D, Jungreuthmayer, C, Zanghellini, J. Metabolomics integrated elementary flux mode analysis in large metabolic networks. Sci Rep. 2015;5:8930–8930. doi:10.1038/srep08930.

263. Pey, J, Planes, FJ. Direct calculation of elementary flux modes satisfying several biological constraints in genome-scale metabolic networks. Bioinformatics. 2014;30:2197–2203. doi:10.1093/bioinformatics/btu193.

264. Kaleta, C, de Figueiredo, LF, Schuster, S. Can the whole be less than the sum of its parts? Pathway analysis in genome-scale metabolic networks using elementary flux patterns. Genome Res. 2009;19:1872–1883. doi:10.1101/gr.090639.108.

265. Jungreuthmayer, C, Ruckerbauer, DE, Zanghellini, J. regEfmtool: Speeding up elementary flux mode calculation using transcriptional regulatory rules in the form of three-state logic. BioSystems. 2013;113:37–39. doi:10.1016/j.biosystems.2013.04.002.

266. Schuster, S, Pfeiffer, T, Moldenhauer, F, Koch, I, Dandekar, T. Exploring the pathway structure of metabolism: decomposition into subnetworks and application to Mycoplasma pneumoniae. Bioinformatics. 2002;18:351–361. doi:10.1093/bioinformatics/18.2.351.

267. Schwartz, JM, Gaugain, C, Nacher, JC, de Daruvar, A, Kanehisa, M. Observing metabolic functions at the genome scale. Genome Biol. 2007;8:R123–R123. doi:10.1186/gb-2007-8-6-r123.

268. Soons, ZI, Ferreira, EC, Rocha, I. Identification of minimal metabolic pathway models consistent with phenotypic data. J Process Control. 2011;21:1483–1492. doi:10.1016/j.jprocont.2011.05.012.

269. Jungers, RM, Zamorano, F, Blondel, VD, Wouwer, AV, Bastin, G. Fast computation of minimal elementary decompositions of metabolic flux vectors. Automatica. 2011;47:1255–1259. doi:10.1016/j.automatica.2011.01.011.

270. Machado, D, Soons, Z, Patil, KR, Ferreira, EC, Rocha, I. Random sampling of elementary flux modes in large-scale metabolic networks. Bioinformatics. 2012;28:i515–i521. doi:10.1093/bioinformatics/bts401.

271. Wishart, DS, Guo, A, Oler, E, Wang, F, Anjum, A, Peters, H, Dizon, R, Sayeeda, Z, Tian, S, Lee, BL, et al. HMDB 5.0: the Human Metabolome Database for 2022. Nucleic Acids Res. 2022;50:D622–D631. doi:10.1093/nar/gkab1062.

272. Moretti, S, Tran, VDT, Mehl, F, Ibberson, M, Pagni, M. MetaNetX/MNXref: unified namespace for metabolites and biochemical reactions in the context of metabolic models. Nucleic Acids Res. 2021;49:D570–D574. doi:10.1093/nar/gkaa992.

273. Preciat Gonzalez, GA, El Assal, LRP, Noronha, A, Thiele, I, Haraldsdóttir, HS, Fleming, RMT. Comparative evaluation of atom mapping algorithms for balanced metabolic reactions: application to Recon 3D. J Cheminform. 2017;9:39–39. doi:10.1186/s13321-017-0223-1.

274. Jeong, S, Eskandari, R, Park, SM, Alvarez, J, Tee, SS, Weissleder, R, Kharas, MG, Lee, H, Keshari, KR. Real-time quantitative analysis of metabolic flux in live cells using a hyperpolarized micromagnetic resonance spectrometer. Sci Adv. 2017;3:e1700341–e1700341. doi:10.1126/sciadv.1700341.

275. Guan, J, Jia, C, Li, Y, Liu, Z, Wang, J, Yang, Z, Gu, C, Su, D, Houk, KN, Zhang, D, et al. Direct single-molecule dynamic detection of chemical reactions. Sci Adv. 2018;4:eaar2177–eaar2177. doi:10.1126/sciadv.aar2177.

276. Hou, S, Exell, J, Welsher, K. Real-time 3D single molecule tracking. Nat Commun. 2020;11:3607–3607. doi:10.1038/s41467-020-17444-6.

277. Li, CY, Duan, S, Yi, J, Wang, C, Radjenovic, PM, Tian, ZQ, Li, JF. Real-time detection of single-molecule reaction by plasmon-enhanced spectroscopy. Sci Adv. 2020;6:eaba6012–eaba6012. doi:10.1126/sciadv.aba6012.

278. Yan, W, Chen, S, Li, P, Dong, R, Shin, HH, Yang, L. Real-time monitoring of a single molecule in sub-nanometer space by dynamic surface-enhanced raman spectroscopy. J Phys Chem Lett. 2023;14:8726–8733. doi:10.1021/acs.jpclett.3c02276.

279. Yang, C, Li, Y, Zhou, S, Guo, Y, Jia, C, Liu, Z, Houk, KN, Dubi, Y, Guo, X. Real-time monitoring of reaction stereochemistry through single-molecule observations of chirality-induced spin selectivity. Nat Chem. 2023;15:972–979. doi:10.1038/s41557-023-01212-2.

280. Anderson, DF, Kurtz, TG. *Stochastic analysis of biochemical systems* (Springer Publishing Company, Incorporated, New York, NY, 2015).

281. Husic, BE, Pande, VS. Markov state models: from an art to a science. J Am Chem Soc. 2018;140:2386–2396. doi:10.1021/jacs.7b12191.

282. Karp, PD, Billington, R, Caspi, R, Fulcher, CA, Latendresse, M, Kothari, A, Keseler, IM, Krummenacker, M, Midford, PE, Ong, Q, et al. The BioCyc collection of microbial genomes and metabolic pathways. Brief Bioinform. 2019;20:1085–1093. doi:10.1093/bib/bbx085.

283. Astero, M, Rousu, J. Enhancing atom mapping with multitask learning and symmetry-aware deep graph matching. J Cheminform. 2025;17:87. doi:10.1186/s13321-025-01030-3.

284. Carlson, R, Fell, D, Srienc, F. Metabolic pathway analysis of a recombinant yeast for rational strain development. Biotechnol Bioeng. 2002;79:121–134. doi:10.1002/bit.10305.

285. Becker, J, Zelder, O, Häfner, S, Schröder, H, Wittmann, C. From zero to hero—design-based systems metabolic engineering of Corynebacterium glutamicum for l-lysine production. Metab Eng. 2011;13:159–168. doi:10.1016/j.ymben.2011.01.003.

286. Papin, JA, Price, ND, Edwards, JS, Palsson, BØ. The genome-scale metabolic extreme pathway structure in Haemophilus influenzae shows significant network redundancy. J Theor Biol. 2002;215:67–82. doi:10.1006/jtbi.2001.2499.

287. Kim, J, Reed, JL. RELATCH: relative optimality in metabolic networks explains robust metabolic and regulatory responses to perturbations. Genome Biol. 2012;13:R78–R78. doi:10.1186/gb-2012-13-9-r78.

288. Yizhak, K, Benyamini, T, Liebermeister, W, Ruppin, E, Shlomi, T. Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. Bioinformatics. 2010;26:i255–i260. doi:10.1093/bioinformatics/btq183.

289. Zhou, J, Zhuang, Y, Xia, J. Integration of enzyme constraints in a genome-scale metabolic model of Aspergillus niger improves phenotype predictions. Microb Cell Fact. 2021;20:125–125. doi:10.1186/s12934-021-01614-2.

290. De Becker, K, Totis, N, Bernaerts, K, Waldherr, S. Using resource constraints derived from genomic and proteomic data in metabolic network models. Curr Opin Syst Biol. 2022;29:100400–100400. doi:10.1016/j.coisb.2021.100400.

291. Bonetta, L. Whole-genome sequencing breaks the cost barrier. Cell. 2010;141:917–919. doi:10.1016/j.cell.2010.05.034.

292. Miggiels, P, Wouters, B, van Westen, GJ, Dubbelman, AC, Hankemeier, T. Novel technologies for metabolomics: more for less. Trends Analyt Chem. 2019;120:115323–115323. doi:10.1016/j.trac.2018.11.021.

293. Poulsen, KM, Pho, T, Champion, JA, Payne, CK. Automation and low-cost proteomics for characterization of the protein corona: experimental methods for big data. Anal Bioanal Chem. 2020;412:6543–6551. doi:10.1007/s00216-020-02726-1.

294. Gygi, SP, Rochon, Y, Franza, BR, Aebersold, R. Correlation between protein and mRNA abundance in yeast. Mol Cell Biol. 1999;19:1720–1730. doi:10.1128/mcb.19.3.1720.

295. Hauf, J, Zimmermann, FK, Müller, S. Simultaneous genomic overexpression of seven glycolytic enzymes in the yeast Saccharomyces cerevisiae. Enzyme Microb Technol. 2000;26:688–698. doi:10.1016/s0141-0229(00)00160-5.

296. Ter Kuile, BH, Westerhoff, HV. Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway. FEBS Lett. 2001;500:169–171. doi:10.1016/S0014-5793(01)02613-8.

297. Ghaemmaghami, S, Huh, WK, Bower, K, Howson, RW, Belle, A, Dephoure, N, O'Shea, EK, Weissman, JS. Global analysis of protein expression in yeast. Nature. 2003;425:737–741. doi:10.1038/nature02046.

298. Greenbaum, D, Colangelo, C, Williams, K, Gerstein, M. Comparing protein abundance and mRNA expression levels on a genomic scale. Genome Biol. 2003;4:117–117. doi:10.1186/gb-2003-4-9-117.

299. Washburn, MP, Koller, A, Oshiro, G, Ulaszek, RR, Plouffe, D, Deciu, C, Winzeler, E, III, JRY. Protein pathway and complex clustering of correlated mRNA and protein expression analyses in Saccharomycescerevisiae. Proc Natl Acad Sci USA. 2003;100:3107–3107. doi:10.1073/pnas.0634629100.

300. Tian, Q, Stepaniants, SB, Mao, M, Weng, L, Feetham, MC, Doyle, MJ, Yi, EC, Dai, H, Thorsson, V, Eng, J, et al. Integrated genomic and proteomic analyses of gene expression in mammalian cells. Mol Cell Proteomics. 2004;3:960–969. doi:10.1074/mcp.M400055-MCP200.

301. Ishihama, Y, Oda, Y, Tabata, T, Sato, T, Nagasu, T, Rappsilber, J, Mann, M. Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. Mol Cell Proteomics. 2005;4:1265–1272. doi:10.1074/mcp.M500061-MCP200.

302. Schmidt, MW, Houseman, A, Ivanov, AR, Wolf, DA. Comparative proteomic and transcriptomic profiling of the fission yeast Schizosaccharomyces pombe. Mol Syst Biol. 2007;3:79–79. doi:10.1038/msb4100117.

303. Nie, L, Wu, G, Zhang, W. Correlation between mRNA and protein abundance in Desulfovibrio vulgaris: a multiple regression to identify sources of variations. Biochem Biophys Res Commun. 2006;339:603–610. doi:10.1016/j.bbrc.2005.11.055.

304. Moxley, J, Jewett, M, Antoniewicz, M, Villas-Boas, S, Alper, H, Wheeler, R, Tong, L, Hinnebusch, A, Ideker, T, Nielsen, JB, et al. Linking transcriptional regulation and high resolution metabolic fluxes in yeast modulated by the global regulator Gcn4p. Proc Natl Acad Sci USA. 2009;106:6477–6482. doi:10.1073/pnas.0811091106.

305. Schwanhäusser, B, Busse, D, Li, N, Dittmar, G, Schuchhardt, J, Wolf, J, Chen, W, Selbach, M. Global quantification of mammalian gene expression control. Nature. 2011;473:337–342. doi:10.1038/nature10098.

306. Edfors, F, Danielsson, F, Hallström, BM, Käll, L, Lundberg, E, Pontén, F, Forsström, B, Uhlén, M. Gene-specific correlation of RNA and protein levels in human cells and tissues. Mol Syst Biol. 2016;12:883–883. doi:10.15252/msb.20167144.

307. Yu, H, Deng, H, He, J, Keasling, JD, Luo, X. UniKP: a unified framework for the prediction of enzyme kinetic parameters. Nat Commun. 2023;14:8211. doi:10.1038/s41467-023-44113-1.

308. Gollub, MG, Backes, T, Kaltenbach, HM, Stelling, J, Mathelier, A. ENKIE: a package for predicting enzyme kinetic parameter values and their uncertainties. Bioinformatics. 2024;40. doi:10.1093/bioinformatics/btae652.

309. Garritano, AN, Song, W, Thomas, T. Carbon fixation pathways across the bacterial and archaeal tree of life. PNAS Nexus. 2022;1:pgac226–pgac226. doi:10.1093/pnasnexus/pgac226.

310. Satanowski, A, Dronsella, B, Noor, E, Vögeli, B, He, H, Wichmann, P, Erb, TJ, Lindner, SN, Bar-Even, A. Awakening a latent carbon fixation cycle in Escherichia coli. Nat Commun. 2020;11:5812–5812. doi:10.1038/s41467-020-19564-5.

311. Sánchez-Andrea, I, Guedes, IA, Hornung, B, Boeren, S, Lawson, CE, Sousa, DZ, Bar-Even, A, Claassens, NJ, Stams, AJM. The reductive glycine pathway allows autotrophic growth of Desulfovibrio desulfuricans. Nat Commun. 2020;11:5090–12. doi:10.1038/s41467-020-18906-7.

312. García, JL, Galán, B. Integrating greenhouse gas capture and C1 biotechnology: a key challenge for circular economy. Microb Biotechnol. 2022;15:228–239. doi:10.1111/1751-7915.13991.

313. Halsey, KH, Carter, AE, Giovannoni, SJ. Synergistic metabolism of a broad range of C1 compounds in the marine methylotrophic bacterium HTCC2181. Environ Microbiol. 2012;14:630–640. doi:10.1111/j.1462-2920.2011.02605.x.

Appendices

A

A Markovian decomposition to uniquely assign elementary flux mode weights in unimolecular flux networks

A.1 Proof of correctness of the CHMC algorithm

In this section, I establish the correctness of the CHMC algorithm described in Chapter 3. By correctness, I mean that it calculates EFM weights that attribute the correct amount of flux for every reaction in the network. The first section below reprises some definitions from Chapter 3. The next section establishes some lemmas about the steady state properties of the Markov chain that tracks a single particle moving around the network. I then establish some lemmas about steady state properties of the CHMC. And in the final section, I complete the proof of the theorem by showing that the fluxes are correctly explained.

A.1.1 Flux network basics

A *flux network* is a triple $G = (V, E, f)$, where $V = \{1, 2, \ldots, n\}$ is a set of nodes (metabolites), $E \subseteq V \times V$ is a set of directed edges (reactions), and f are positive edge weights (fluxes). I assume no node has an edge directly back to itself: $(i, i) \notin E$ for all $i \in V$. f may be defined either as a function $f : E \to \mathbb{R}$, or as a function $f : V \times V \to \Re^{\geq 0}$ where $f_{i,j} > 0 \iff (i, j) \in E$.

I restrict attention to strongly-connected flux networks, meaning that every node is reachable from every other. I also restrict attention to steady-state flux networks, meaning that the flux into any node equals the flux out of the node: for all $i \in V$, $\sum_{j \neq i} f_{i,j} = \sum_{j \neq i} f_{j,i}$. I call this the flux *at* node i, and overload the f notation to denote the flux at node i as f_i.

I define the *total network flux* as $f_{tot} = \sum_{i,j} f_{i,j}$, and I define the *normalized fluxes* as $f^n_{i,j} = f_{i,j}/f_{tot}$. Obviously, if a set of elementary flux mode (EFM) weights explains all the normalized fluxes f^n in a network, then f_{tot} times those weights explains all

unnormalized fluxes f.

Finally, as already implied by the definitions above, I restrict attention to closed flux networks, meaning there is no flux into or out of the system. However, this is a matter merely of mathematical convenience. Any open flux network, where sources nodes produce flux and sink nodes consume flux, can be trivially handled by our approach as well. One simply represents the "exterior" of the network by a special node, and connects flux providing and consuming edges to the source and sink nodes respectively.

In closed flux networks, the EFMs are the simple cycles of in network–paths that begin and end at the same node, with no nodes repeated in between. I use $C = \{c_1, c_2, \ldots, c_k\}$ to denote this set of cycles. Any particular cycle is of the form$(v_1, v_2, \ldots, v_{m-1}, v_m)$, where $v_1 = v_m \in V$, but no other v_i are equal to each other.

A.1.2 Markov chain steady state properties

Based on the flux network G I define a Markov chain $M = (S, T, s_0)$ where the state space is the same as the nodes of the flux network: $S = V$. The state-to-state transition probabilities T are proportional to the fluxes in the following sense: where s_t denotes a random variable representing the state of the chain at time t, $T_{i,j} \equiv Pr(s_{t+1} = j | s_t = i) = f_{i,j} / \sum_{j'} f_{i,j'}$. I arbitrarily choose the initial state of the Markov chain to be $s_0 = 1$ since the choice does not impact the long-term elementary flux mode (EFM) probabilities.

Because the chain network G is strongly connected, the Markov chain M is irreducible. It therefore has a unique steady state distribution π satisfying $\pi \cdot T = \pi$. Indeed, the steady state probabilities are just the normalized fluxes at the nodes:

Lemma 1. *The steady state distribution π of the Markov chain M is equal to the vector of normalized state fluxes f^n of the flux network G: $\pi = f^n$.*

Proof: Let f^n be a row vector of the fluxes at the states $1, \ldots, n$, and consider the j^{th} column of the Markov transition matrix, which I will denote $T_{\cdot,j}$. Then:

$$f^n \cdot T_{\cdot,j} = \sum_i f_i^n \cdot T_{i,j} = \sum_i f_i^n \cdot \frac{f_{i,j}}{f_i^n} = \sum_i f_{i,j}^n = f_j^n$$

Since this is true for any j, then I have $f^n \cdot T = f^n$. Note also that $\sum_i f^n = \sum_{i,j} f_{i,j}^n = \sum_{i,j} f_{i,j}/f_{tot} = f_{tot}/f_{tot} = 1$. Since π is the unique steady state vector satisfying $\pi \cdot T = \pi$ and $\sum_i \pi = 1$, and since f^n satisfies the same two properties, then $\pi = f^n$.

I define the probabilistic flux $\pi_{i,j}$ of the Markov chain as the steady state frequency with which transitions from state i to state j occur, or equivalently, $\pi_{i,j} = \pi_i \cdot T_{i,j}$. It follows immediately that $\pi_{i,j} = f_{i,j}^n$, which I state in the lemma below without further argument.

Lemma 2. *The probabilistic fluxes $\pi_{i,j}$ of the Markov chain M are equal to the normalized fluxes $f_{i,j}^n$ of the flux network G.*

A.1.3 Cycle-History Markov chain steady state properties

The cycle-history Markov chain (CHMC) is a Markov chain $M^c = (S^c, T^c, s_0^c)$ that is constructed based on the Markov chain M. Every state $s \in S^c$ in the CHMC corresponds to a simple path of chain M starting from its initial state s_0. That is, $s \equiv (s_0, s_1, \ldots, s_m)$ for some $m \geq 0$, where all $s_i \in S$, $s_0 = 1$, and $s_i \neq s_j$ for all $0 \leq i < j \leq m$. The state set S^c corresponds to all possible simple paths in M, where "possible paths" means paths with strictly positive probability. The initial state of the CHMC corresponds to the trivial zero-step path starting at M's initial state: $s_0^c \equiv (s_0)$. For every CHMC state $s \equiv (s_0, s_1, \ldots, s_m)$ and for every Markov chain state j that can be transitioned to from s_m (i.e. every j where $T_{s_m,j} > 0$), there is a corresponding transition in the CHMC with the same probability. If state j is not on

the path $(s_0, s_1, \ldots, s_m)$, then the transition is to a CHMC state $s' \equiv (s_0, s_1, \ldots, s_m, j)$. If state j is on the path $(s_0, s_1, \ldots, s_m)$, say at $s_k = j$, then it is a transition to the CHMC state corresponding to the shorter path $s' \equiv (s_0, s_1, \ldots, s_k)$. This latter type of transition is called a cycle-/loop-/EFM-closing transition, because it corresponds to the Markov chain finishing transiting a cycle—specifically, the cycle given by the latter portion of s's path: $(s_k, s_{k+1}, \ldots, s_m, s_k)$. In the CHMC, it is self-evident that every state is reachable from every other state. Therefore, the CHMC has a unique steady state distribution π^c satisfying $\pi^c \cdot T^c = \pi^c$.

Whereas the CHMC construction "blows up" the Markov chain M's state space into a much larger state space of all possible paths, it is useful to consider a sort of reverse transformation where I aggregate together different states of the CHMC based on the final state of their path. That is, define a function $A : S^c \mapsto S$ where for $s \equiv (s_0, s_1, \ldots, s_m)$, I have $A(s) = s_m$. Recalling that the state set of the Markov chain M, and the original flux network G, are just the numbers $1, 2, \ldots, n$, then A is just a mapping from S^c to the numbers 1 to n. For convenience, let us also define the set of CHMC states mapping to each Markov state i: $S_i^c = \{s \in S^c : A(s) = i\}$.

I define the aggregated CHMC steady state distribution as π^A via: $\pi_i^A = \sum_{s \in S_i^c} \pi_s^c$. In words, π^A is a vector where the i^{th} element is the sum of steady state probabilities of all CHMC state corresponding to paths ending at i. Our first main observation about the CHMC is that the aggregated steady state probabilities are equal to the Markov chain steady states.

Lemma 3. *Where π is the steady state distribution of the Markov chain M, and where π^A are the aggregated steady state probabilities of the CHMC M^c: $\pi^A = \pi$.*

This makes intuitive sense. The CHMC is basically keeping multiple "copies" of a Markov state i based on the path followed to reach state i. However, i always has to be reached by some path. Therefore, by summing over all of those paths, I should

arrive at the total probability of being in state i. However, I can argue more formally as follows.

Proof: Consider any Markov state i and consider any CHMC state corresponding to a path ending at that state, $s \in S_i^c$. Let $T^c_{\cdot,s}$ denote the column of the CHMC transition matrix corresponding to transitions into state s. Then I have

$$\pi_s^c = \pi^c \cdot T^c_{\cdot,s} = \sum_{s' \in S^c} \pi^c_{s'} \cdot T^c_{s',s}$$

The summation over s' can alternatively be written as a double sum over Markov chain states j and CHMC states corresponding to paths ending at j:

$$\pi_s^c = \sum_{j \in S} \sum_{s' \in S_j^c} \pi^c_{s'} \cdot T^c_{s',s}$$

I can further sum both sides of the equation over all CHMC states s'' in the same partition as s:

$$\sum_{s'' \in S_i^c} \pi^c_{s''} = \sum_{s'' \in S_i^c} \sum_{j \in S} \sum_{s' \in S_j^c} \pi^c_{s'} \cdot T^c_{s',s''}$$

I recognize the left hand side as the aggregated CHMC probability of states corresponding to paths ending at state i. Meanwhile, the right hand side can be rewritten by reordering the sums.

$$\pi_i^A = \sum_{j \in S} \sum_{s' \in S_j^c} \sum_{s'' \in S_i^c} \pi^c_{s'} \cdot T^c_{s',s''}$$

The double summation over s' and s'' is not really a double summation, in the sense that for any $s' \in S_j^c$, and assuming a transition to some $s'' \in S_i^c$ is possible at all (i.e. assuming $T_{j,i} > 0$) then there will be one and precisely one s'' to which s' can transition. All other transitions from s' to other elements of S_i^c will have probability zero. To be precise, if s'' is either a shorter part of the path represented by s', or if s'' is a one-step extension of the path represented by s', then I will have $T^c_{s',s''} = T_{j,i}$. Otherwise, I will

have $T^c_{s',s''} = 0$. Therefore, I can rewrite as:

$$\pi^A_i = \sum_{j \in S} \sum_{s' \in S^c_j} \pi^c_{s'} \cdot T_{j,i}$$

From this, I deduce that:

$$\pi^A_i = \sum_{j \in S} \pi^A_j \cdot T_{j,i}$$

Because this is true for any Markov state i, I must have $\pi^A = \pi^A \cdot T$, meaning that π^A is the (unique) steady state distribution for the Markov chain M, or in other words $\pi = \pi^A$, establishing the lemma.

A similar result is true for the aggregated probabilistic fluxes of the CHMC. The probabilistic flux from CHMC state s to state s' is $\pi^c_{s,s'} = \pi^c_s \cdot T^c_{s,s'}$. I then define the aggregated probabilistic fluxes for any $i, j \in S$, as $\pi^A_{i,j} = \sum_{s \in S^c_i} \sum_{s' \in S^c_j} \pi^c_{s,s'}$.

Lemma 4. *The aggregated probabilistic fluxes $\pi^A_{i,j}$ of the CHMC M^c are equal to the probabilistic fluxes $\pi_{i,j}$ of the Markov chain M, which are equal to the normalized fluxes $f^n_{i,j}$ of the flux network G.*

Proof: I start from the definition of the aggregated probabilistic flux:

$$\pi^A_{i,j} = \sum_{s \in S^c_i} \sum_{s' \in S^c_j} \pi^c_{s,s'} = \sum_{s \in S^c_i} \sum_{s' \in S^c_j} \pi^c_s \cdot T^c_{s,s'} = \sum_{s \in S^c_i} \pi^c_s \cdot \sum_{s' \in S^c_j} T^c_{s,s'}$$

As explained in the previous section, the apparent double sum over s and s' really only has one non-zero term per s, because there is only one (at most) non-zero transition from any element of S^c_i to any element of S^c_j. Therefore, I can write:

$$\pi^A_{i,j} = \sum_{s \in S^c_i} \pi^c_s \cdot T_{i,j} = \pi^A_i \cdot T_{i,j} = \pi_i \cdot T_{i,j} = \pi_{i,j}$$

which completes the lemma.

A.1.4 Explaining fluxes

I begin by establishing a means by which CHMC probabilistic fluxes $\pi^c_{s,s'}$ (not the aggregated ones) can be "explained" in terms of weighted cycles, and then I extend that result to show that a simple aggregation and rescaling of those same weights explains the fluxes $f_{i,j}$ in the original flux network G.

I first show that the CHMC probabilistic fluxes can be explained by interpreting them as a flux network. Specifically, consider the flux network $G' = (V', E', f')$ where the nodes are the same as the CHMC states: $V' = S^c$. The transitions are $(s, s') \in E'$ iff $T^c_{s,s'} > 0$. And the fluxes are $f' = \pi^c$. Because this is a flux network, I know that the fluxes can be explained by a weighted combination of elementary flux modes. What are the EFMs of G'? First, G' is a closed network, and so the EFMs are a set of cycles, $C' = \{c'_1, c'_2, \ldots, c'_{k'}\}$. However, it is a very special set of cycles. By construction the CHMC transitions form a tree, with the exception of the loop-closing transition. Therefore, every cycle c' contains precisely one loop-closure edge. And conversely, no other cycle c'' includes that same loop-closure edge. Therefore, let us denote those loop closure edges as $(s_i, s'_i) \in E'$ for $1 \leq i \leq k'$, one for each cycle. When an EFM is the one and only EFM to pass through a certain edge in a flux network, then the flux (or weight) of that EFM must equal that flux of that edge—because there is no other way to explain the flux on that edge. And since every EFM has one such edge, I know therefore that there can only be one weighting of the EFMs that explains the probabilistic flux f', namely: EFM i must be given weight $f_{s_i,s'_i} = \pi^c_{s_i,s'_i}$. Therefore, I have the following lemma.

Lemma 5. *There is a set of non-negative weights $w'_1, w'_2, \ldots, w'_{k'}$ that fully explain (probabilistic) fluxes f' in the sense that: For all $s, s' \in V'$:*

$$f'_{s,s'} = \sum_{i=1}^{k'} w'_i \cdot \begin{cases} 1 & \text{if } (s, s') \in c'_i \\ 0 & \text{otherwise.} \end{cases}$$

In particular, the weights are $w'_i = \pi^c_{s_i,s'_i}$ *where* (s_i, s'_i) *is the loop-closure edge in the* i^{th} *cycle (a.k.a. EFM)* c'_i.

Finally, I connect the weights of the previous lemma to the original flux network G. As stated in the main manuscript, by construction of the Markov chain M, the CHMC M^c, and the flux network G', every EFM in G' corresponds to an EFM in G. However, as also demonstrated in the main text, an EFM in G can correspond to more than one EFM in G'. In essence, G' can have multiple "copies" of the same EFM in G, prefaced by different paths reaching that EFM. Therefore, where $C = \{c_1, c_2, \ldots, c_k\}$ are the EFMs of G and $C' = \{c'_1, c'_2, \ldots, c'_{k'}\}$ are the EFMs of G, I can define a many-to-one mapping relating the two: $Z : C' \mapsto C$. And let us define the aggregated weights w_i^Z for $1 \leq i \leq k$ as: $w_i^Z = \sum_{\{c'_j : Z(c'_j) = c_i\}} w'_j$, where w'_j are the EFM weights for G' of the previous lemma. And finally, let us define the scaled aggregated weights as $w_1, \ldots, w_k$ where $w_i = f_{tot} \cdot w_i^Z$. Then I arrive at the main theorem.

Theorem 2. *The weights* $w_1, w_2, \ldots, w_k$ *for the EFMs* C *of flux network* G *fully explain the normalized fluxes* f *in the sense that: For all* $i, j \in V$*:*

$$f_{i,j} = \sum_{l=1}^{k} w_l \cdot \begin{cases} 1 & \text{if } (i,j) \in c_l \\ 0 & \text{otherwise.} \end{cases}$$

Proof: Consider any $i, j \in V$. Starting from the right hand side of the equation in

the theorem, I have:

$$\begin{aligned}
&\sum_{l=1}^{k} w_l \cdot \begin{cases} 1 & \text{if } (i,j) \in c_l \\ 0 & \text{otherwise.} \end{cases} \\
= &\sum_{l=1}^{k} f_{tot} \cdot w_l^Z \cdot \begin{cases} 1 & \text{if } (i,j) \in c_l \\ 0 & \text{otherwise.} \end{cases} \\
= &\sum_{l=1}^{k} f_{tot} \cdot \left(\sum_{c'_j : Z(c'_j) = c_l} w'_j \right) \cdot \begin{cases} 1 & \text{if } (i,j) \in c_l \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

The double sum is, equivalently, a sum over all cycles c' of the CHMC M^c (or network G'). However, the 0/1 condition should be 1 only for those cycles containing what amounts to an i to j transition in the original Markov chain M:

$$\begin{aligned}
= &\sum_{m=1}^{k'} f_{tot} \cdot w'_m \cdot \begin{cases} 1 & \text{if } \exists s \in S_i^c, s' \in S_j^c \text{ } s.t. \text{ } (s,s') \in c'_m \\ 0 & \text{otherwise.} \end{cases} \\
= &f_{tot} \sum_{m=1}^{k'} w'_m \sum_{s \in S_i^c, s' \in S_j^c} \begin{cases} 1 & \text{if } (s,s') \in c'_m \\ 0 & \text{otherwise.} \end{cases} \\
= &f_{tot} \sum_{s \in S_i^c, s' \in S_j^c} \sum_{m=1}^{k'} w'_m \begin{cases} 1 & \text{if } (s,s') \in c'_m \\ 0 & \text{otherwise.} \end{cases} \\
= &f_{tot} \sum_{s \in S_i^c, s' \in S_j^c} f'_{s,s'}
\end{aligned}$$

where the last step follows from Lemma 5. Then:

$$\begin{aligned}
&= f_{tot} \sum_{s \in S_i^c, s' \in S_j^c} \pi^c_{s,s'} \\
&= f_{tot} \cdot \pi^A_{i,j} \\
&= f_{tot} \cdot \pi_{i,j}
\end{aligned}$$

where the last step comes from Lemma 4. Then by Lemma 2 I have:

$$\begin{aligned} &= f_{tot} \cdot f_{i,j}^{n} \\ &= f_{tot} \cdot \frac{f_{i,j}}{f_{tot}} \\ &= f_{i,j} \end{aligned}$$

And so the theorem is proved.

A.2 List of optimization-based method solvers

Table A.1: List of solvers for each optimization-based method. All except for Gurobi are open-source and did not require a manual installation step. See `https://jump.dev/JuMP.jl/stable/installation/#Supported-solvers` for more details.

	COSMO	OSQP	Gurobi	SCIP	CDDLib	ECOS	GLPK	ProxSDP	Tulip
Minimize L2	✓	✓							
Maximize qSPA	✓	✓							
Maximize lSPA		✓	✓	✓	✓	✓	✓	✓	✓
Minimize lSPA		✓	✓	✓	✓	✓	✓	✓	✓
Minimize milAP			✓						

A.3 Sphingolipid kinetic model (reproduced with minor corrections)

Table A.2: Healthy and Alzheimer's disease reaction parameters in the sphingolipid kinetic model. Inhibition parameters were not provided by Wronowska *et al.* and were manually estimated to satisfy flux conservation and steady state lipid concentrations in the healthy model. ($Ki_{S1P} = 0.0356898872566651$ and $Ki_{C1P} = 0.1784494362833260$).

#	Reaction	Flux equation	Healthy			Alzheimers		
			Km (nmol/mg)	Vm (nmol/min/mg)	k (1/min)	Km (nmol/mg)	Vm (nmol/min/mg)	r (1/min)
1	→ ER.CER	$\frac{Vm_1}{(1+C.S1P/Ki_{S1P})\times(1+GA.C1P/Ki_{C1P})}$	0	–	–	0	–	
2	ER.CER → ER.SPH	$\frac{Vm_2\times ER.CER}{Km_2+ER.CER}$	0.08100	0.48000	–	0.08100	0.32000	–
3	ER.SPH → ER.CER	$\frac{Vm_3\times ER.SPH}{Km_3+ER.SPH}$	0.17100	2.40000	–	0.17100	2.40000	–
4	ER.SPH → ER.S1P	$\frac{Vm_4\times ER.SPH}{Km_4+ER.SPH}$	0.00340	0.17500	–	0.00340	0.87500	–
5	ER.S1P → ER.SPH	$\frac{Vm_5\times ER.S1P}{Km_5+ER.S1P}$	0.03850	3.10000	–	0.03850	3.10000	–
6	ER.S1P → ER.PhET + 2THD	$\frac{Vm_5\times ER.S1P}{Km_5+ER.S1P}$	0.03500	0	–	0.03500	0	–
7	ER.CER → N.CER	$k_7\times ER.CER - r_7\times N.CER$	–	–	0.80000	–	–	0.80000
8	N.CER → N.SM	$\frac{Vm_8\times N.CER}{Km_8+N.CER}$	0.15500	0.10900	–	0.15500	0.10900	–
9	N.SM → N.CER	$\frac{Vm_9\times N.SM}{Km_9+N.SM}$	0.12600	0.00113	–	0.12600	0.10000	–
10	N.CER → N.SPH	$\frac{Vm_{10}\times N.CER}{Km_{10}+N.CER}$	0.06010	0.06800	–	0.06010	0.00453	–
11	N.SPH → N.S1P	$\frac{Vm_{11}\times N.SPH}{Km_{11}+N.SPH}$	0.00340	0.10000	–	0.00340	0.05000	–
12	N.S1P → N.SPH	$\frac{Vm_{12}\times N.S1P}{Km_{12}+N.S1P}$	0.02500	1.24000	–	0.02500	1.24000	–
13	N.SM → ER.SM	$k_{13}\times N.SM - r_{13}\times ER.SM$	–	–	0.12000	–	–	0.12000
14	N.SPH → C.SPH	$k_{14}\times N.SPH - r_{14}\times C.SPH$	–	–	0.50000	–	–	0.50000
15	N.S1P → C.S1P	$k_{15}\times N.S1P - r_{15}\times C.S1P$	–	–	0.44000	–	–	0.44000
16	C.S1P → M.S1P	$k_{16}\times C.S1P - r_{16}\times M.S1P$	–	–	0.25000	–	–	0.25000
17	M.S1P → M.SPH	$\frac{Vm_{17}\times M.S1P}{Km_{17}+M.S1P}$	0.03850	3.00000	–	0.03850	3.00000	–
18	M.SPH → M.S1P	$\frac{Vm_{18}\times M.SPH}{Km_{18}+M.SPH}$	0.00340	0.15000	–	0.00340	0.07000	–
19	M.SPH → M.CER	$\frac{Vm_{19}\times M.SPH}{Km_{19}+M.SPH}$	0.00250	0.10000	–	0.00250	0.10000	–
20	M.CER → M.SPH	$\frac{Vm_{20}\times M.CER}{Km_{20}+M.CER}$	0.14900	2.27000	–	0.14900	2.00000	–
21	M.SPH → C.SPH	$k_{21}\times M.SPH - r_{21}\times C.SPH$	–	–	0.43000	–	–	0.43000
22	C.SPH → ER.SPH	$k_{22}\times C.SPH - r_{22}\times ER.SPH$	–	–	23.0000	–	–	23.0000
23	ER.SM → IM.SM	$k_{23}\times ER.SM - r_{23}\times IM.SM$	–	–	0.01000	–	–	0.01000
24	ER.SM → OM.SM	$k_{24}\times ER.SM - r_{24}\times OM.SM$	–	–	0.00600	–	–	0.01500
25	OM.SM → OM.CER	$\frac{Vm_{25}\times M.CER}{Km_{20}+M.CER}$	0.04550	0.00100	–	0.04550	0.00200	–
26	OM.CER → OM.SPH	$\frac{Vm_{26}\times OM.CER}{Km_{26}+OM.CER}$	0.06010	0.12000	–	0.06010	0.08000	–
27	OM.SPH → OM.S1P	$\frac{Vm_{27}\times OM.SPH}{Km_{27}+OM.SPH}$	0.03400	0.11000	–	0.03400	0.05500	–
28	IM.SM → IM.CER	$\frac{Vm_{28}\times IM.SM}{Km_{28}+IM.SM}$	0.14800	0.00200	–	0.14800	0.00250	–
29	IM.CER → IM.SPH	$\frac{Vm_{29}\times IM.CER}{Km_{29}+IM.CER}$	0.06010	0.03000	–	0.06010	0.02000	–
30	IM.SPH → IM.S1P	$\frac{Vm_{30}\times IM.SPH}{Km_{30}+IM.SPH}$	0.00506	0.00300	–	0.00506	0.00150	–
31	OM.CER → IM.CER	$k_{31}\times OM.CER - r_{31}\times IM.CER$	–	–	1.00000	–	–	1.00000
32	OM.SPH → IM.SPH	$k_{32}\times OM.SPH - r_{32}\times IM.SPH$	–	–	1.00000	–	–	1.00000
33	IM.S1P → OM.S1P	$k_{33}\times IM.S1P - r_{33}\times OM.S1P$	–	–	1.00000	–	–	1.00000
34	IM.S1P → C.S1P	$k_{34}\times IM.S1P - r_{34}\times C.S1P$	–	–	0.40000	–	–	0.40000
35	IM.SPH → C.SPH	$k_{35}\times IM.SPH - r_{35}\times C.SPH$	–	–	3.00000	–	–	3.00000
36	C.SPH → C.S1P	$\frac{Vm_{36}\times C.SPH}{Km_{36}+C.SPH}$	0.00340	0.00360	–	0.00340	0.00100	–
37	C.S1P → ER.S1P	$k_{37}\times C.S1P - r_{37}\times ER.S1P$	–	–	4.50000	–	–	4.50000
38	ER.CER → M.CER	$k_{38}\times ER.CER - r_{38}\times M.CER$	–	–	1.00000	–	–	1.00000
39	ER.CER → GA.CER	$k_{39}\times ER.CER - r_{39}\times GA.CER$	–	–	5.00000	–	–	1.00000

Table A.2 continued.

#	Reaction	Flux equation	Healthy			Alzheimers		
			Km (nmol/mg)	Vm (nmol/min/mg)	k (1/min)	Km (nmol/mg)	Vm (nmol/min/mg)	r (1/min)
40	ER.CER → GA.CER	$k_{40} \times ER.CER$	–	–	0.08000	–	–	0.02000
41	GACF.CER → GACF.GluCER	$\frac{Vm_{41} \times GACF.CER}{Km_{41}+GACF.CER}$	0.04000	0.01000	–	0.40000	0.01000	–
42	GACF.GluCER → GA.GluCER	$k_{42} \times GACF.GluCER$	–	–	1.00000	–	–	1.00000
43	GA.GluCER → GA.LacCER	$\frac{Vm_{43} \times GA.GluCER}{Km_{43}+GA.GluCER}$	0.00300	0.00100	–	0.00300	0.00100	–
44	GA.LacCER → GA.GSL	$\frac{Vm_{44} \times GA.LacCER}{Km_{44}+GA.LacCER}$	0.00300	0.00100	–	0.00300	0.00100	–
45	GA.GSL → OM.GSL	$k_{45} \times GA.GSL$	–	–	0.20000	–	–	0.20000
46	OM.GSL → L.GSL	$k_{46} \times OM.GSL$	–	–	0.03000	–	–	0.03000
47	L.GSL → L.CER	$\frac{Vm_{47} \times GA.LacCER}{Km_{44}+GA.LacCER}$	0.01900	0.00610	–	0.01900	0.00610	–
48	OM.SM → L.SM	$k_{48} \times OM.SM$	–	–	0.00200	–	–	0.01100
49	L.SM → L.CER	f_1	0.04550	0.00183	–	0.04550	0.01000	–
50	L.CER → L.SPH	$\frac{Vm_{50} \times L.CER}{Km_{50}+L.CER}$	0.14900	0.10000	–	0.14900	0.00669	–
51	L.SPH → C.SPH	$k_{51} \times L.SPH$	–	–	1.00000	–	–	1.00000
52	GA.SPH → C.SPH	$k_{52} \times GA.SPH - r_{52} \times C.SPH$	–	–	0.20000	–	–	0.20000
53	GA.SPH → GA.S1P	$\frac{Vm_{53} \times GA.SPH}{Km_{53}+GA.SPH}$	0.00506	0.13000	–	0.00506	0.08000	–
54	GA.S1P → GA.SPH	$\frac{Vm_{54} \times GA.S1P}{Km_{54}+GA.S1P}$	0.03600	2.00000	–	0.03600	2.00000	–
55	GA.CER → GA.SM	$\frac{Vm_{55} \times GA.CER}{Km_{55}+GA.CER}$	0.02000	0.30000	–	0.02000	0.30000	–
56	GA.SM → OM.SM	$k_{56} \times GA.SM$	–	–	0.01500	–	–	0.04500
57	GA.SM → GA.CER	$\frac{Vm_{57} \times GA.SM}{Km_{57}+GA.SM}$	0.14800	0.05000	–	0.14800	0.07000	–
58	GA.CER → GA.SPH	$\frac{Vm_{58} \times GA.CER}{Km_{58}+GA.CER}$	0.08100	0.80000	–	0.08100	0.52900	–
59	GA.CER → GA.C1P	$\frac{Vm_{59} \times GA.CER}{Km_{59}+GA.CER}$	0.10700	2.00000	–	0.10700	5.00000	–
60	GA.C1P → GA.CER	$\frac{Vm_{60} \times GA.C1P}{Km_{60}+GA.C1P}$	0.03600	0.75000	–	0.03600	0.75000	–
61	GA.C1P → OM.C1P	$k_{61} \times GA.C1P$	–	–	2.50000	–	–	2.50000
62	OM.C1P → OM.CER	$\frac{Vm_{62} \times OM.C1P}{Km_{62}+OM.C1P}$	0.03600	0.78000	–	0.03600	0.78000	–
63	OM.CER → OM.C1P	$\frac{Vm_{63} \times OM.CER}{Km_{63}+OM.CER}$	0.10700	2.10000	–	0.10700	0.50000	–
64	→ OM.SM	k_{64}	–	–	0	–	–	0
65	→ OM.C1P	$k_{65} - r_{65} \times OM.C1P$	–	–	0	–	–	0
66	→ OM.CER	k_{66}	–	–	0	–	–	0
67	→ OM.SPH	$k_{67} - r_{67} \times OM.SPH$	–	–	0	–	–	0
68	→ OM.S1P	$k_{68} - r_{68} \times OM.S1P$	–	–	0	–	–	0
69	OM.S1P → OM.SPH	$\frac{Vm_{69} \times OM.S1P}{Km_{69}+OM.S1P}$	0.03600	0.43000	–	0.03600	0.43000	–

Note: $f_1 = \frac{Vm_{49} \times L.SM}{(Km_{49}+L.SM) \times (1+GA.C1P/Ki_{C1P}) \times (1+C.S1P/Ki_{S1P})}$

Table A.3: System of ordinary differential equations of the sphingolipid network.

$$\frac{\mathrm{d}[GA.C1P]}{\mathrm{d}t} = \frac{Vm_{59} \times [GA.CER]}{Km_{59} + [GA.CER]} - \frac{Vm_{60} \times [GA.C1P]}{Km_{60} + [GA.C1P]} - k_{61} \times [GA.C1P] \tag{A.1}$$

$$\frac{\mathrm{d}[OM.C1P]}{\mathrm{d}t} = k_{61} \times [GA.C1P] - \frac{Vm_{62} \times [OM.C1P]}{Km_{62} + [OM.C1P]} + \frac{Vm_{63} \times [OM.CER]}{Km_{63} + [OM.CER]} + k_{65} - r_{65} \times [OM.C1P] \tag{A.2}$$

$$\begin{aligned}\frac{\mathrm{d}[ER.CER]}{\mathrm{d}t} = {} & \frac{Vm_1}{\left(\frac{1+[C.S1P]}{Ki_{S1P}} \times \frac{1+[GA.C1P]}{Ki_{C1P}}\right)} - \frac{Vm_2 \times [ER.CER]}{Km_2 + [ER.CER]} + \frac{Vm_3 \times [ER.SPH]}{Km_3 + [ER.SPH]} \\ & - (k_7 \times [ER.CER] - r_7 \times [N.CER]) - (k_{38} \times [ER.CER] \\ & - r_{38} \times [M.CER]) - (k_{39} \times [ER.CER]) - (k_{40} \times [ER.CER])\end{aligned} \tag{A.3}$$

$$\frac{\mathrm{d}[IM.CER]}{\mathrm{d}t} = \frac{Vm_{28} \times [IM.SM]}{Km_{28} + [IM.SM]} - \frac{Vm_{29} \times [IM.CER]}{Km_{29} + [IM.CER]} - (k_{31} \times [IM.CER] - r_{31} \times [OM.CER]) \tag{A.4}$$

$$\begin{aligned}\frac{\mathrm{d}[L.CER]}{\mathrm{d}t} = {} & \frac{Vm_{47} \times [L.GSL]}{Km_{47} + [L.GSL]} + \frac{Vm_{49} \times [L.SM]}{(Km_{49} + [L.SM]) \times (1 + [GA.C1P]/Ki_{C1P}) \times (1 + [C.S1P]/Ki_{S1P})} \\ & - \frac{Vm_{50} \times [L.CER]}{Km_{50} + [L.CER]}\end{aligned} \tag{A.5}$$

$$\begin{aligned}\frac{\mathrm{d}[GA.CER]}{\mathrm{d}t} = {} & (k_{39} \times [ER.CER]) - \frac{Vm_{55} \times [GA.CER]}{Km_{55} + [GA.CER]} + \frac{Vm_{57} \times [GA.SM]}{Km_{57} + [GA.SM]} - \frac{Vm_{58} \times [GA.CER]}{Km_{58} + [GA.CER]} \\ & - \frac{Vm_{59} \times [GA.CER]}{Km_{59} + [GA.CER]} + \frac{Vm_{60} \times [GA.C1P]}{Km_{60} + [GA.C1P]}\end{aligned} \tag{A.6}$$

$$\frac{\mathrm{d}[GACF.CER]}{\mathrm{d}t} = (k_{40} \times [ER.CER]) - \frac{Vm_{41} \times [GACF.CER]}{Km_{41} + [GACF.CER]} \tag{A.7}$$

$$\frac{\mathrm{d}[M.CER]}{\mathrm{d}t} = \frac{Vm_{19} \times [M.SPH]}{Km_{19} + [M.SPH]} - \frac{Vm_{20} \times [M.CER]}{Km_{20} + [M.CER]} + (k_{38} \times [ER.CER] - r_{38} \times [M.CER]) \tag{A.8}$$

$$\frac{\mathrm{d}[N.CER]}{\mathrm{d}t} = (k_7 \times [ER.CER] - r_7 \times [N.CER]) - \frac{Vm_8 \times [N.CER]}{Km_8 + [N.CER]} + \frac{Vm_9 \times [N.SM]}{Km_9 + [N.SM]} - \frac{Vm_{10} \times [N.CER]}{Km_{10} + [N.CER]} \tag{A.9}$$

$$\begin{aligned}\frac{\mathrm{d}[OM.CER]}{\mathrm{d}t} = {} & \frac{Vm_{25} \times [OM.SM]}{Km_{25} + [OM.SM]} - \frac{Vm_{26} \times [OM.CER]}{Km_{26} + [OM.CER]} + (k_{31} \times [IM.CER] - r_{31} \times [OM.CER]) \\ & + \frac{Vm_{62} \times [OM.C1P]}{Km_{62} + [OM.C1P]} - \frac{Vm_{63} \times [OM.CER]}{Km_{63} + [OM.CER]} + k_{66}\end{aligned} \tag{A.10}$$

$$\frac{\mathrm{d}[GA.GluCER]}{\mathrm{d}t} = (k_{42} \times [GACF.GluCER]) - \frac{Vm_{43} \times [GA.GluCER]}{Km_{43} + [GA.GluCER]} \tag{A.11}$$

$$\frac{\mathrm{d}[GACF.GluCER]}{\mathrm{d}t} = \frac{Vm_{41} \times [GACF.CER]}{Km_{41} + [GACF.CER]} - (k_{42} \times [GACF.GluCER]) \tag{A.12}$$

$$\frac{\mathrm{d}[L.GSL]}{\mathrm{d}t} = (k_{46} \times [OM.GSL]) - \frac{Vm_{47} \times [L.GSL]}{Km_{47} + [L.GSL]} \tag{A.13}$$

$$\frac{\mathrm{d}[GA.GSL]}{\mathrm{d}t} = \frac{Vm_{44} \times [GA.LacCER]}{Km_{44} + [GA.LacCER]} - (k_{45} \times [GA.GSL]) \tag{A.14}$$

$$\frac{\mathrm{d}[OM.GSL]}{\mathrm{d}t} = (k_{45} \times [GA.GSL]) - (k_{46} \times [OM.GSL]) \tag{A.15}$$

$$\frac{\mathrm{d}[GA.LacCER]}{\mathrm{d}t} = \frac{Vm_{43} \times [GA.GluCER]}{Km_{43} + [GA.GluCER]} - \frac{Vm_{44} \times [GA.LacCER]}{Km_{44} + [GA.LacCER]} \tag{A.16}$$

$$\begin{aligned}\frac{\mathrm{d}[C.S1P]}{\mathrm{d}t} = {} & (k_{15} \times [N.S1P] - r_{15} \times [C.S1P]) - (k_{16} \times [C.S1P] - r_{16} \times [M.S1P]) + (k_{34} \times [IM.S1P] \\ & - r_{34} \times [C.S1P]) + \frac{Vm_{36} \times [C.SPH]}{Km_{36} + [C.SPH]} - (k_{37} \times [C.S1P] - r_{37} \times [ER.S1P])\end{aligned} \tag{A.17}$$

$$\frac{\mathrm{d}[ER.S1P]}{\mathrm{d}t} = \frac{Vm_4 \times [ER.SPH]}{Km_4 + [ER.SPH]} - \frac{Vm_5 \times [ER.S1P]}{Km_5 + [ER.S1P]} - \frac{Vm_6 \times [ER.S1P]}{Km_6 + [ER.S1P]} + (k_{37} \times [C.S1P] - r_{37} \times [ER.S1P]) \tag{A.18}$$

$$\frac{\mathrm{d}[IM.S1P]}{\mathrm{d}t} = \frac{Vm_{30} \times [IM.SPH]}{Km_{30} + [IM.SPH]} - (k_{33} \times [IM.S1P]) - (k_{34} \times [IM.S1P] - r_{34} \times [C.S1P]) \tag{A.19}$$

$$\frac{\mathrm{d}[GA.S1P]}{\mathrm{d}t} = \frac{Vm_{53} \times [GA.SPH]}{Km_{53} + [GA.SPH]} - \frac{Vm_{54} \times [GA.S1P]}{Km_{54} + [GA.S1P]} \tag{A.20}$$

$$\frac{\mathrm{d}[M.S1P]}{\mathrm{d}t} = (k_{16} \times [C.S1P] - r_{16} \times [M.S1P]) - \frac{Vm_{17} \times [M.S1P]}{Km_{17} + [M.S1P]} + \frac{Vm_{18} \times [M.SPH]}{Km_{18} + [M.SPH]} \tag{A.21}$$

Table A.3 continued.

$$\frac{\mathrm{d}[N.S1P]}{\mathrm{d}t} = \frac{Vm_{11} \times [N.SPH]}{Km_{11} + [N.SPH]} - \frac{Vm_{12} \times [N.S1P]}{Km_{12} + [N.S1P]} - (k_{15} \times [N.S1P] - r_{15} \times [C.S1P]) \tag{A.22}$$

$$\frac{\mathrm{d}[OM.S1P]}{\mathrm{d}t} = \frac{Vm_{27} \times [OM.SPH]}{Km_{27} + [OM.SPH]} + (k_{33} \times [IM.S1P]) + (k_{68} - r_{68} \times [OM.S1P]) - \frac{Vm_{69} \times [OM.S1P]}{Km_{69} + [OM.S1P]} \tag{A.23}$$

$$\begin{aligned}\frac{\mathrm{d}[ER.SM]}{\mathrm{d}t} &= (k_{13} \times [N.SM] - r_{13} \times [ER.SM]) - (k_{23} \times [ER.SM] - r_{23} \times [IM.SM]) - (k_{24} \times [ER.SM] \\ &\quad - r_{24} \times [OM.SM])\end{aligned} \tag{A.24}$$

$$\frac{\mathrm{d}[IM.SM]}{\mathrm{d}t} = (k_{23} \times [ER.SM] - r_{23} \times [IM.SM]) - \frac{Vm_{28} \times [IM.SM]}{Km_{28} + [IM.SM]} \tag{A.25}$$

$$\frac{\mathrm{d}[L.SM]}{\mathrm{d}t} = (k_{48} \times [OM.SM]) - \frac{Vm_{49} \times [L.SM]}{(Km_{49} + [L.SM]) \times (1 + [GA.C1P]/Ki_{C1P}) \times (1 + [C.S1P]/Ki_{S1P})} \tag{A.26}$$

$$\frac{\mathrm{d}[GA.SM]}{\mathrm{d}t} = \frac{Vm_{55} \times [GA.CER]}{Km_{55} + [GA.CER]} - (k_{56} \times [GA.SM]) - \frac{Vm_{57} \times [GA.SM]}{Km_{57} + [GA.SM]} \tag{A.27}$$

$$\frac{\mathrm{d}[N.SM]}{\mathrm{d}t} = \frac{Vm_{8} \times [N.CER]}{Km_{8} + [N.CER]} - \frac{Vm_{9} \times [N.SM]}{Km_{9} + [N.SM]} - (k_{13} \times [N.SM] - r_{13} \times [ER.SM]) \tag{A.28}$$

$$\frac{\mathrm{d}[OM.SM]}{\mathrm{d}t} = (k_{24} \times [ER.SM] - r_{24} \times [OM.SM]) - \frac{Vm_{25} \times [OM.SM]}{Km_{25} + [OM.SM]} - (k_{48} \times [OM.SM]) + (k_{56} \times [GA.SM]) + (k_{64}) \tag{A.29}$$

$$\begin{aligned}\frac{\mathrm{d}[C.SPH]}{\mathrm{d}t} &= (k_{14} \times [N.SPH] - r_{14} \times [C.SPH]) + (k_{21} \times [M.SPH] - r_{21} \times [C.SPH]) \\ &\quad - (k_{22} \times [C.SPH] - r_{22} \times [ER.SPH]) + (k_{35} \times [IM.SPH] - r_{35} \times [C.SPH]) \\ &\quad - \frac{Vm_{36} \times [C.SPH]}{Km_{36} + [C.SPH]} + (k_{51} \times [L.SPH]) - (k_{52} \times [C.SPH] - r_{52} \times [GA.SPH])\end{aligned} \tag{A.30}$$

$$\begin{aligned}\frac{\mathrm{d}[ER.SPH]}{\mathrm{d}t} &= \frac{Vm_{2} \times [ER.CER]}{Km_{2} + [ER.CER]} - \frac{Vm_{3} \times [ER.SPH]}{Km_{3} + [ER.SPH]} - \frac{Vm_{4} \times [ER.SPH]}{Km_{4} + [ER.SPH]} + \frac{Vm_{5} \times [ER.S1P]}{Km_{5} + [ER.S1P]} \\ &\quad + (k_{22} \times [C.SPH] - r_{22} \times [ER.SPH])\end{aligned} \tag{A.31}$$

$$\begin{aligned}\frac{\mathrm{d}[IM.SPH]}{\mathrm{d}t} &= \frac{Vm_{29} \times [IM.CER]}{Km_{29} + [IM.CER]} - \frac{Vm_{30} \times [IM.SPH]}{Km_{30} + [IM.SPH]} - (k_{32} \times [IM.SPH] - r_{32} \times [OM.SPH]) \\ &\quad - (k_{35} \times [IM.SPH] - r_{35} \times [C.SPH])\end{aligned} \tag{A.32}$$

$$\frac{\mathrm{d}[L.SPH]}{\mathrm{d}t} = \frac{Vm_{50} \times [L.CER]}{Km_{50} + [L.CER]} - (k_{51} \times [L.SPH]) \tag{A.33}$$

$$\frac{\mathrm{d}[GA.SPH]}{\mathrm{d}t} = (k_{52} \times [C.SPH] - r_{52} \times [GA.SPH]) - \frac{Vm_{53} \times [GA.SPH]}{Km_{53} + [GA.SPH]} + \frac{Vm_{54} \times [GA.S1P]}{Km_{54} + [GA.S1P]} + \frac{Vm_{58} \times [GA.CER]}{Km_{58} + [GA.CER]} \tag{A.34}$$

$$\begin{aligned}\frac{\mathrm{d}[M.SPH]}{\mathrm{d}t} &= \frac{Vm_{17} \times [M.S1P]}{Km_{17} + [M.S1P]} - \frac{Vm_{18} \times [M.SPH]}{Km_{18} + [M.SPH]} - \frac{Vm_{19} \times [M.SPH]}{Km_{19} + [M.SPH]} + \frac{Vm_{20} \times [M.CER]}{Km_{20} + [M.CER]} \\ &\quad - (k_{21} \times [M.SPH] - r_{21} \times [C.SPH])\end{aligned} \tag{A.35}$$

$$\frac{\mathrm{d}[N.SPH]}{\mathrm{d}t} = \frac{Vm_{10} \times [N.CER]}{Km_{10} + [N.CER]} - \frac{Vm_{11} \times [N.SPH]}{Km_{11} + [N.SPH]} + \frac{Vm_{12} \times [N.S1P]}{Km_{12} + [N.S1P]} - (k_{14} \times [N.SPH] - r_{14} \times [C.SPH]) \tag{A.36}$$

$$\begin{aligned}\frac{\mathrm{d}[OM.SPH]}{\mathrm{d}t} &= \frac{Vm_{26} \times [OM.CER]}{Km_{26} + [OM.CER]} - \frac{Vm_{27} \times [OM.SPH]}{Km_{27} + [OM.SPH]} + (k_{32} \times [IM.SPH] - r_{32} \times [OM.SPH]) \\ &\quad + (k_{67} - r_{67} \times [OM.SPH]) + \frac{Vm_{69} \times [OM.S1P]}{Km_{69} + [OM.S1P]}\end{aligned} \tag{A.37}$$

Table A.4: Steady state sphingolipid concentrations for the healthy and Alzheimer's disease kinetic model.

#	Class	Healthy concentration (nmol/mg)	Alzheimer's disease concentration (nmol/mg)
1	GA.C1P	0.001459	0.002544
2	OM.C1P	0.002373	0.001674
3	ER.CER	0.004556	0.015994
4	IM.CER	0.002204	0.006029
5	L.CER	0.002455	0.191720
6	GA.CER	0.001788	0.001209
7	GACF.CER	0.001513	0.001322
8	M.CER	0.004094	0.007883
9	N.CER	0.001980	0.057990
10	OM.CER	0.002322	0.006417
11	GA.GluCER	0.001720	0.001411
12	GACF.GluCER	0.000364	0.000320
13	L.GSL	0.001207	0.001051
14	GA.GSL	0.001822	0.001599
15	OM.GSL	0.012148	0.010663
16	GA.LacCER	0.001720	0.001411
17	C.S1P	0.000327	0.000154
18	ER.S1P	0.001218	0.008694
19	IM.S1P	0.000620	0.000438
20	GA.S1P	0.000663	0.000218
21	M.S1P	0.001021	0.000801
22	N.S1P	0.001135	0.000592
23	OM.S1P	0.000645	0.000503
24	ER.SM	0.160374	0.243261
25	IM.SM	0.132108	0.199413
26	L.SM	0.105423	0.024589
27	GA.SM	0.123667	0.041210
28	N.SM	0.011979	0.041789
29	OM.SM	0.628298	0.313096
30	C.SPH	0.001162	0.001432
31	ER.SPH	0.003909	0.006367
32	IM.SPH	0.001796	0.003148
33	L.SPH	0.001621	0.003764
34	GA.SPH	0.001945	0.000896
35	M.SPH	0.003631	0.023567
36	N.SPH	0.004038	0.004664
37	OM.SPH	0.002293	0.003768

Table A.5: Steady state sphingolipid fluxes for the healthy and Alzheimer's disease kinetic model.

#	Reaction	Healthy flux (nmol/mg/min)	Alzheimer's disease flux (nmol/mg/min)
1	→ ER.CER	0	0
2	ER.CER → ER.SPH	0.025560	0.052766
3	ER.SPH → ER.CER	0.053634	0.086157
4	ER.SPH → ER.S1P	0.093591	0.570411
5	ER.S1P → ER.SPH	0.095056	0.571062
6	ER.S1P → ER.PhET + 2THD	0	0
7	ER.CER → N.CER	0.003446	0.006996
8	N.CER → N.SM	0.001375	0.029677
9	N.SM → N.CER	0.000098	0.024906
10	N.CER → N.SPH	0.002169	0.002224
11	N.SPH → N.S1P	0.054287	0.028918
12	N.S1P → N.SPH	0.053869	0.028696
13	N.SM → ER.SM	0.001277	0.004771
14	N.SPH → C.SPH	0.001752	0.002002
15	N.S1P → C.S1P	0.000418	0.000222
16	C.S1P → M.S1P	0.000037	0.000003
17	M.S1P → M.SPH	0.077500	0.061178
18	M.SPH → M.S1P	0.077463	0.061174
19	M.SPH → M.CER	0.059223	0.090409
20	M.CER → M.SPH	0.060708	0.100491
21	M.SPH → C.SPH	0.001522	0.010085
22	C.SPH → ER.SPH	0.026610	0.032739
23	ER.SM → IM.SM	0.000943	0.001436
24	ER.SM → OM.SM	0.000334	0.003336
25	OM.SM → OM.CER	0.000932	0.001746
26	OM.CER → OM.SPH	0.004463	0.007718
27	OM.SPH → OM.S1P	0.006950	0.005487
28	IM.SM → IM.CER	0.000943	0.001435
29	IM.CER → IM.SPH	0.001061	0.001823
30	IM.SPH → IM.S1P	0.000786	0.000575
31	OM.CER → IM.CER	0.000118	0.000388
32	OM.SPH → IM.SPH	0.005083	0.008156
33	IM.S1P → OM.S1P	0.000620	0.000438
34	IM.S1P → C.S1P	0.000166	0.000137
35	IM.SPH → C.SPH	0.005358	0.009404
36	C.SPH → C.S1P	0.000917	0.000296
37	C.S1P → ER.S1P	0.001464	0.000652
38	ER.CER → M.CER	0.001485	0.010082
39	ER.CER → GA.CER	0.022779	0.015994
40	ER.CER → GA.CER	0.000364	0.000320

Table A.5 continued.

#	Reaction	Healthy flux (nmol/mg/min)	Alzheimer's disease flux (nmol/mg/min)
41	GACF.CER → GACF.GluCER	0.000364	0.000320
42	GACF.GluCER → GA.GluCER	0.000364	0.000320
43	GA.GluCER → GA.LacCER	0.000364	0.000320
44	GA.LacCER → GA.GSL	0.000364	0.000320
45	GA.GSL → OM.GSL	0.000364	0.000320
46	OM.GSL → L.GSL	0.000364	0.000320
47	L.GSL → L.CER	0.000364	0.000320
48	OM.SM → L.SM	0.001256	0.003444
49	L.SM → L.CER	0.001256	0.003444
50	L.CER → L.SPH	0.001621	0.003764
51	L.SPH → C.SPH	0.001621	0.003764
52	GA.SPH → C.SPH	0.017275	0.007779
53	GA.SPH → GA.S1P	0.036100	0.012037
54	GA.S1P → GA.SPH	0.036100	0.012037
55	GA.CER → GA.SM	0.024616	0.017100
56	GA.SM → OM.SM	0.001855	0.001854
57	GA.SM → GA.CER	0.022761	0.015246
58	GA.CER → GA.SPH	0.017275	0.007779
59	GA.CER → GA.C1P	0.032867	0.055861
60	GA.C1P → GA.CER	0.029218	0.049501
61	GA.C1P → OM.C1P	0.003648	0.006360
62	OM.C1P → OM.CER	0.048244	0.034650
63	OM.CER → OM.C1P	0.044595	0.028290
64	→ OM.SM	0	0
65	→ OM.C1P	0	0
66	→ OM.CER	0	0
67	→ OM.SPH	0	0
68	→ OM.S1P	0	0
69	OM.S1P → OM.SPH	0.007570	0.005926

A.4 Sphingolipid network elementary flux modes

Table A.6: Description of each EFM in Table A.7.

EFM	Metabolite sequence
1	GA.SM → GA.CER → GA.SM
2	ER.CER → GA.CER → GA.SM → OM.SM → OM.CER → IM.CER → IM.SPH → IM.S1P → C.S1P → M.S1P → M.SPH → C.SPH → ER.SPH → ER.CER
3	ER.CER → ER.SPH → ER.CER
4	M.SPH → C.SPH → ER.SPH → ER.CER → N.CER → N.SM → ER.SM → OM.SM → OM.CER → OM.SPH → IM.SPH → IM.S1P → C.S1P → M.S1P → M.SPH
5	M.SPH → M.S1P → M.SPH
6	N.S1P → N.SPH → N.S1P
7	N.S1P → C.S1P → ER.S1P → ER.SPH → ER.CER → N.CER → N.SPH → N.S1P
8	GA.S1P → GA.SPH → GA.S1P
9	M.SPH → C.SPH → C.S1P → M.S1P → M.SPH
10	L.SPH → C.SPH → ER.SPH → ER.CER → N.CER → N.SM → ER.SM → OM.SM → L.SM → L.CER → L.SPH
11	N.SM → N.CER → N.SM
12	M.SPH → M.CER → M.SPH
13	M.SPH → C.SPH → ER.SPH → ER.CER → M.CER → M.SPH
14	C.SPH → ER.SPH → ER.CER → GACF.CER → GACF.GluCER → GA.GluCER → GA.LacCER → GA.GSL → OM.GSL → L.GSL → L.CER → L.SPH → C.SPH
15	C.SPH → C.S1P → ER.S1P → ER.SPH → ER.CER → GACF.CER → GACF.GluCER → GA.GluCER → GA.LacCER → GA.GSL → OM.GSL → L.GSL → L.CER → L.SPH → C.SPH
16	IM.CER → IM.SPH → C.SPH → ER.SPH → ER.CER → N.CER → N.SM → ER.SM → IM.SM → IM.CER
17	OM.SPH → IM.SPH → IM.S1P → OM.S1P → OM.SPH
18	OM.SPH → OM.S1P → OM.SPH
19	IM.SM → IM.CER → IM.SPH → C.SPH → C.S1P → ER.S1P → ER.SPH → ER.CER → N.CER → N.SM → ER.SM → IM.SM
20	GA.SPH → C.SPH → ER.SPH → ER.CER → GA.CER → GA.SPH
21	IM.SM → IM.CER → IM.SPH → IM.S1P → C.S1P → M.S1P → M.SPH → C.SPH → ER.SPH → ER.CER → N.CER → N.SM → ER.SM → IM.SM
22	OM.SM → OM.CER → IM.CER → IM.SPH → IM.S1P → C.S1P → ER.S1P → ER.SPH → ER.CER → GA.CER → GA.SM → OM.SM

Table A.6 continued.

EFM	Metabolite sequence
23	N.SPH → C.SPH → C.S1P → ER.S1P → ER.SPH → ER.CER → N.CER → N.SPH
24	IM.SM → IM.CER → IM.SPH → IM.S1P → C.S1P → ER.S1P → ER.SPH → ER.CER → N.CER → N.SM → ER.SM → IM.SM
25	ER.CER → GA.CER → GA.SM → OM.SM → L.SM → L.CER → L.SPH → C.SPH → ER.SPH → ER.CER
26	ER.SM → OM.SM → L.SM → L.CER → L.SPH → C.SPH → C.S1P → ER.S1P → ER.SPH → ER.CER → N.CER → N.SM → ER.SM
27	M.SPH → C.SPH → C.S1P → ER.S1P → ER.SPH → ER.CER → M.CER → M.SPH
28	OM.SM → OM.CER → OM.SPH → IM.SPH → IM.S1P → C.S1P → ER.S1P → ER.SPH → ER.CER → GA.CER → GA.SM → OM.SM
29	OM.SM → OM.CER → IM.CER → IM.SPH → IM.S1P → C.S1P → ER.S1P → ER.SPH → ER.CER → N.CER → N.SM → ER.SM → OM.SM
30	ER.S1P → ER.SPH → ER.S1P
31	OM.SM → OM.CER → OM.SPH → IM.SPH → IM.S1P → C.S1P → ER.S1P → ER.SPH → ER.CER → N.CER → N.SM → ER.SM → OM.SM
32	ER.CER → GA.CER → GA.SM → OM.SM → OM.CER → OM.SPH → IM.SPH → C.SPH → ER.SPH → ER.CER
33	ER.SM → OM.SM → OM.CER → OM.SPH → IM.SPH → C.SPH → C.S1P → ER.S1P → ER.SPH → ER.CER → N.CER → N.SM → ER.SM
34	N.S1P → C.S1P → M.S1P → M.SPH → C.SPH → ER.SPH → ER.CER → N.CER → N.SPH → N.S1P
35	N.SPH → C.SPH → ER.SPH → ER.CER → N.CER → N.SPH
36	GA.CER → GA.C1P → OM.C1P → OM.CER → IM.CER → IM.SPH → C.SPH → C.S1P → ER.S1P → ER.SPH → ER.CER → GA.CER
37	GA.CER → GA.C1P → OM.C1P → OM.CER → OM.SPH → IM.SPH → C.SPH → C.S1P → ER.S1P → ER.SPH → ER.CER → GA.CER
38	IM.SPH → C.SPH → ER.SPH → ER.CER → N.CER → N.SM → ER.SM → OM.SM → OM.CER → OM.SPH → IM.SPH
39	OM.CER → OM.C1P → OM.CER
40	L.SPH → C.SPH → C.S1P → ER.S1P → ER.SPH → ER.CER → GA.CER → GA.SM → OM.SM → L.SM → L.CER → L.SPH
41	GA.SPH → C.SPH → C.S1P → ER.S1P → ER.SPH → ER.CER → GA.CER → GA.SPH
42	ER.CER → GA.CER → GA.SM → OM.SM → OM.CER → OM.SPH → IM.SPH → IM.S1P → C.S1P → M.S1P → M.SPH → C.SPH → ER.SPH → ER.CER
43	ER.CER → GA.CER → GA.SM → OM.SM → OM.CER → IM.CER → IM.SPH → C.SPH → C.S1P → ER.S1P → ER.SPH → ER.CER

Table A.6 continued.

EFM	Metabolite sequence
44	GA.CER → GA.C1P → OM.C1P → OM.CER → IM.CER → IM.SPH → C.SPH → ER.SPH → ER.CER → GA.CER
45	ER.SM → OM.SM → OM.CER → IM.CER → IM.SPH → C.SPH → ER.SPH → ER.CER → N.CER → N.SM → ER.SM
46	GA.CER → GA.C1P → OM.C1P → OM.CER → OM.SPH → IM.SPH → C.SPH → ER.SPH → ER.CER → GA.CER
47	M.SPH → C.SPH → ER.SPH → ER.CER → N.CER → N.SM → ER.SM → OM.SM → OM.CER → IM.CER → IM.SPH → IM.S1P → C.S1P → M.S1P → M.SPH
48	IM.SPH → C.SPH → C.S1P → ER.S1P → ER.SPH → ER.CER → N.CER → N.SM → ER.SM → OM.SM → OM.CER → IM.CER → IM.SPH
49	OM.SM → OM.CER → OM.SPH → IM.SPH → C.SPH → C.S1P → ER.S1P → ER.SPH → ER.CER → GA.CER → GA.SM → OM.SM
50	GA.CER → GA.C1P → OM.C1P → OM.CER → IM.CER → IM.SPH → IM.S1P → C.S1P → ER.S1P → ER.SPH → ER.CER → GA.CER
51	GA.CER → GA.C1P → OM.C1P → OM.CER → OM.SPH → IM.SPH → IM.S1P → C.S1P → ER.S1P → ER.SPH → ER.CER → GA.CER
52	GA.CER → GA.C1P → OM.C1P → OM.CER → IM.CER → IM.SPH → IM.S1P → C.S1P → M.S1P → M.SPH → C.SPH → ER.SPH → ER.CER → GA.CER
53	ER.CER → GA.CER → GA.SM → OM.SM → OM.CER → IM.CER → IM.SPH → C.SPH → ER.SPH → ER.CER
54	GA.CER → GA.C1P → OM.C1P → OM.CER → OM.SPH → IM.SPH → IM.S1P → C.S1P → M.S1P → M.SPH → C.SPH → ER.SPH → ER.CER → GA.CER
55	GA.CER → GA.C1P → GA.CER

A.5 Markov weights for the healthy and Alzheimer's disease sphingolipid network

Table A.7: EFM weights for the Markov method in the healthy and Alzheimer's disease condition. Column name c refers to the number of reactions across unique compartments in the sphingolipid network (9 total). Column name l refers to the number of reactions in each EFM.

EFM	w_{wt}	w_{ad}	$\log_2\left(\frac{w_{ad}}{w_{wt}}\right)$	$w_{ad} - w_{wt}$	c	l
1	2.276×10^{-2}	1.525×10^{-2}	-5.781×10^{-1}	-7.515×10^{-3}	1	2
2	1.449×10^{-8}	2.176×10^{-9}	-2.736	-1.232×10^{-8}	6	13
3	2.556×10^{-2}	5.277×10^{-2}	1.046	2.721×10^{-2}	1	2
4	9.882×10^{-8}	7.779×10^{-8}	-3.452×10^{-1}	-2.103×10^{-8}	6	14
5	7.746×10^{-2}	6.117×10^{-2}	-3.406×10^{-1}	-1.629×10^{-2}	1	2
6	5.387×10^{-2}	2.870×10^{-2}	-9.086×10^{-1}	-2.517×10^{-2}	1	2
7	4.080×10^{-4}	2.208×10^{-4}	-8.855×10^{-1}	-1.871×10^{-4}	3	7
8	3.610×10^{-2}	1.204×10^{-2}	-1.585	-2.406×10^{-2}	1	2
9	2.293×10^{-5}	1.535×10^{-6}	-3.901	-2.139×10^{-5}	2	4
10	1.855×10^{-4}	2.194×10^{-3}	3.564	2.008×10^{-3}	5	10
11	9.810×10^{-5}	2.491×10^{-2}	7.988	2.481×10^{-2}	1	2
12	5.922×10^{-2}	9.041×10^{-2}	6.103×10^{-1}	3.119×10^{-2}	1	2
13	1.437×10^{-3}	9.992×10^{-3}	2.798	8.555×10^{-3}	3	5
14	3.526×10^{-4}	3.170×10^{-4}	-1.535×10^{-1}	-3.559×10^{-5}	6	12
15	1.185×10^{-5}	2.855×10^{-6}	-2.054	-8.999×10^{-6}	6	14
16	8.850×10^{-4}	1.402×10^{-3}	6.640×10^{-1}	5.173×10^{-4}	4	9
17	6.198×10^{-4}	4.385×10^{-4}	-4.991×10^{-1}	-1.813×10^{-4}	2	4
18	6.950×10^{-3}	5.487×10^{-3}	-3.409×10^{-1}	-1.463×10^{-3}	1	2
19	2.975×10^{-5}	1.263×10^{-5}	-1.236	-1.712×10^{-5}	4	11
20	1.671×10^{-2}	7.710×10^{-3}	-1.116	-9.004×10^{-3}	3	5
21	6.725×10^{-7}	1.045×10^{-7}	-2.686	-5.680×10^{-7}	5	13
22	5.972×10^{-7}	4.263×10^{-7}	-4.863×10^{-1}	-1.709×10^{-7}	5	11
23	5.697×10^{-5}	1.787×10^{-5}	-1.672	-3.909×10^{-5}	3	7
24	2.771×10^{-5}	2.047×10^{-5}	-4.365×10^{-1}	-7.234×10^{-6}	4	11
25	1.030×10^{-3}	1.220×10^{-3}	2.434×10^{-1}	1.893×10^{-4}	5	9
26	6.234×10^{-6}	1.975×10^{-5}	1.664	1.352×10^{-5}	5	12
27	4.830×10^{-5}	8.997×10^{-5}	8.975×10^{-1}	4.167×10^{-5}	3	7
28	2.262×10^{-5}	8.472×10^{-6}	-1.417	-1.414×10^{-5}	5	11
29	1.075×10^{-7}	7.668×10^{-7}	2.835	6.593×10^{-7}	5	12
30	9.359×10^{-2}	5.704×10^{-1}	2.608	4.768×10^{-1}	1	2
31	4.071×10^{-6}	1.524×10^{-5}	1.904	1.117×10^{-5}	5	12
32	7.224×10^{-4}	5.803×10^{-4}	-3.160×10^{-1}	-1.421×10^{-4}	5	9
33	4.372×10^{-6}	9.399×10^{-6}	1.104	5.028×10^{-6}	5	12
34	9.903×10^{-6}	1.127×10^{-6}	-3.135	-8.775×10^{-6}	4	9
35	1.695×10^{-3}	1.985×10^{-3}	2.280×10^{-1}	2.901×10^{-4}	3	5
36	2.961×10^{-6}	2.680×10^{-6}	-1.436×10^{-1}	-2.805×10^{-7}	5	11
37	1.121×10^{-4}	5.326×10^{-5}	-1.074	-5.886×10^{-5}	5	11
38	1.300×10^{-4}	1.044×10^{-3}	3.005	9.138×10^{-4}	5	10
39	4.459×10^{-2}	2.829×10^{-2}	-6.566×10^{-1}	-1.631×10^{-2}	1	2
40	3.463×10^{-5}	1.098×10^{-5}	-1.657	-2.365×10^{-5}	5	11

Table A.7 continued.

EFM	w_{wt}	w_{ad}	$\log_2\left(\frac{w_{ad}}{w_{wt}}\right)$	$w_{ad} - w_{wt}$	c	l
41	5.619×10^{-4}	6.942×10^{-5}	-3.017	-4.924×10^{-4}	3	7
42	5.490×10^{-7}	4.324×10^{-8}	-3.666	-5.057×10^{-7}	6	13
43	6.412×10^{-7}	2.629×10^{-7}	-1.286	-3.783×10^{-7}	5	11
44	8.807×10^{-5}	2.976×10^{-4}	1.757	2.096×10^{-4}	5	9
45	3.434×10^{-6}	5.252×10^{-5}	3.935	4.909×10^{-5}	5	10
46	3.335×10^{-3}	5.915×10^{-3}	8.266×10^{-1}	2.580×10^{-3}	5	9
47	2.609×10^{-9}	3.914×10^{-9}	5.850×10^{-1}	1.305×10^{-9}	6	14
48	1.154×10^{-7}	4.730×10^{-7}	2.035	3.575×10^{-7}	5	12
49	2.428×10^{-5}	5.225×10^{-6}	-2.216	-1.906×10^{-5}	5	11
50	2.757×10^{-6}	4.345×10^{-6}	6.563×10^{-1}	1.588×10^{-6}	5	11
51	1.044×10^{-4}	8.636×10^{-5}	-2.740×10^{-1}	-1.806×10^{-5}	5	11
52	6.692×10^{-8}	2.218×10^{-8}	-1.593	-4.474×10^{-8}	6	13
53	1.907×10^{-5}	2.920×10^{-5}	6.142×10^{-1}	1.012×10^{-5}	5	9
54	2.535×10^{-6}	4.408×10^{-7}	-2.523	-2.094×10^{-6}	6	13
55	2.922×10^{-2}	4.950×10^{-2}	7.606×10^{-1}	2.028×10^{-2}	1	2

A.6 Individual flux reconstruction error across methods

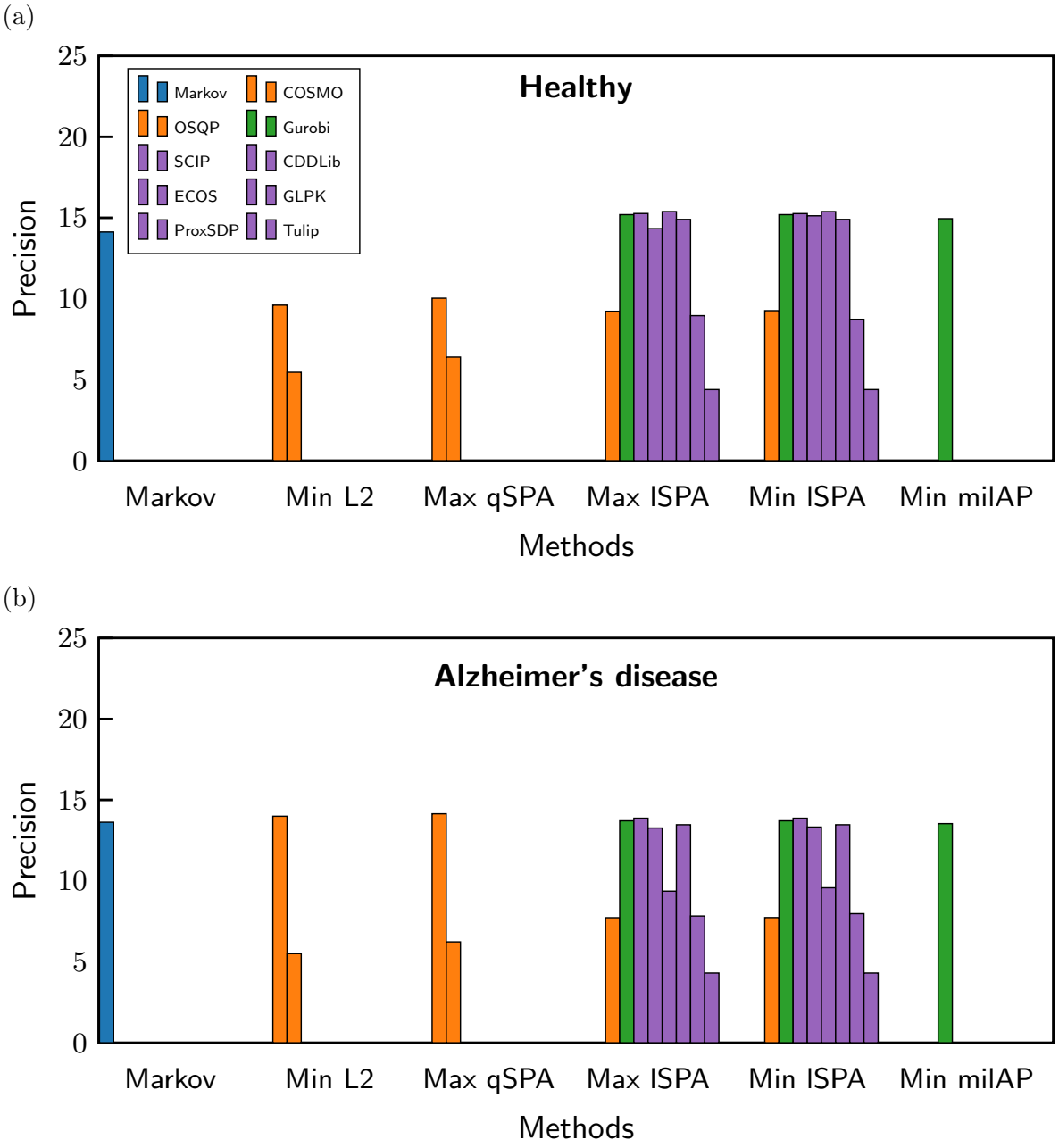


Figure A.1: Error between individual fluxes reconstructed from EFM weights from different methods and the observed network fluxes for (a) healthy sphingolipid network and (b) Alzheimer's disease sphingolipid network. Precision measured as $-\log_{10}(\sum (Aw - v)^2/|v|)$ where A is the binary matrix of EFM weights (rows are reactions, columns are EFMs), w is the set of EFM weights, and v is the vector of observed network fluxes.

A.7 Total flux reconstruction error across methods

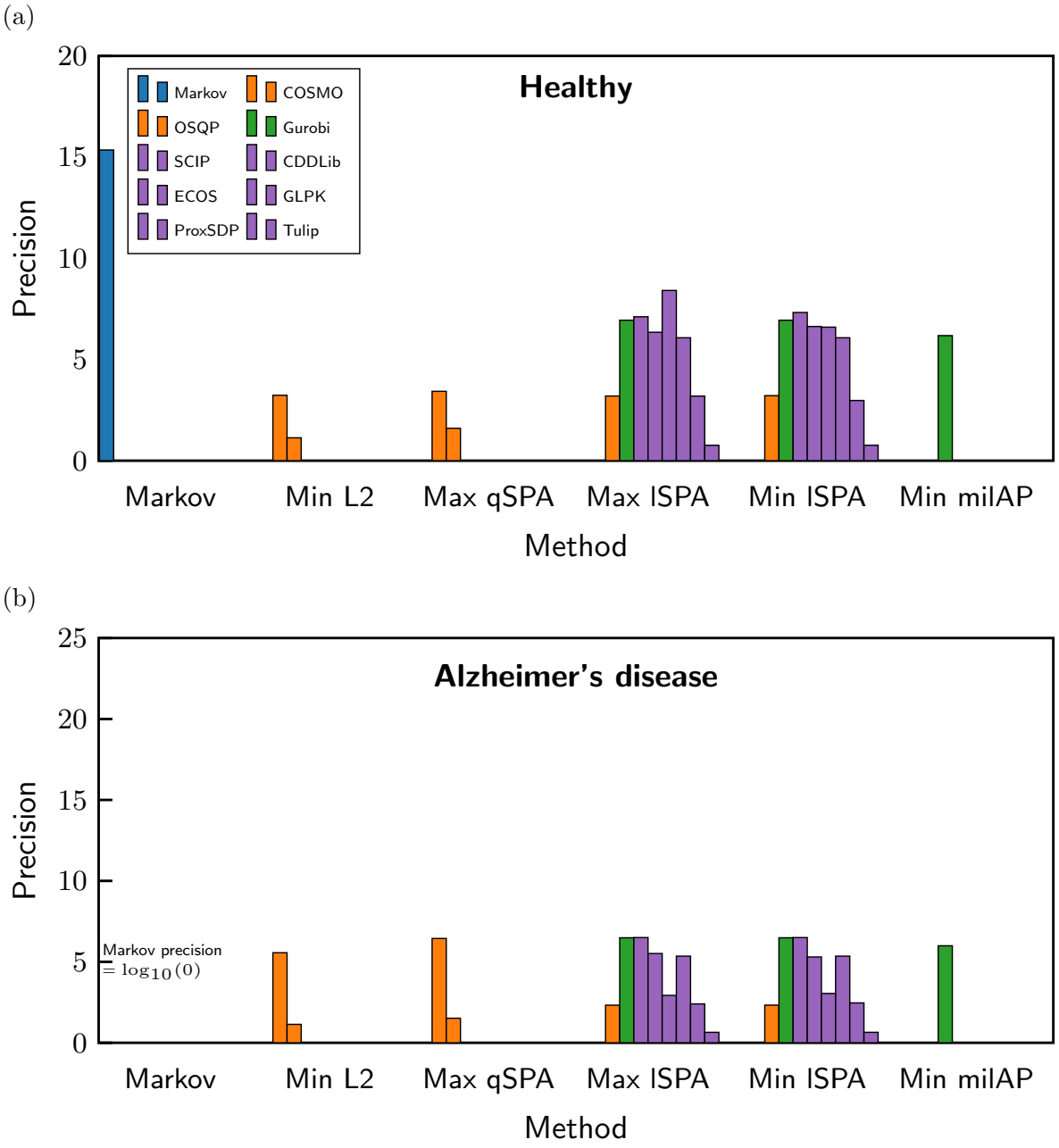


Figure A.2: Error between total fluxes reconstructed from EFM weights from different methods and the observed network fluxes for (a) healthy sphingolipid network and (b) Alzheimer's disease sphingolipid network. Precision measured as $-\log_{10}(\sum (Aw - v)^2/|v|)$ where A is the binary matrix of EFM weights (rows are reactions, columns are EFMs), w is the set of EFM weights, and v is the vector of observed network fluxes.

B

A dominant pathway flux hypothesis

B.1 Solving for the power law distribution parameter k by maximum likelihood estimation

The power law distribution for the continuous random variable x is defined as:

$$f(x) = a(x_{\min}, k)x^{-k} \tag{B.1}$$

where k is the power law exponent, and a is the normalizing constant that depends on the value of k for some value of $x \geq x_{\min}$. Parameters such as k can be solved by maximum likelihood estimation (MLE) which maximizes the (log-)likelihood function of the distribution conditioned on the observed data. This can be achieved in the

`fitdistrplus` package in R for many types of distributions. However, this package does not support the power law distribution since the likelihood function includes $x_{\min}$ and increases as $x_{\min}$ grows larger. This motivates an analytical solution to fit power laws to the steady state flux and rescaled elementary flux mode (EFM) weight data.

Starting with Equation (B.1), the likelihood function of the power law is written as

$$L(x_i|x_{\min}, k) = \prod_{i=1}^{N} f(x_i|x_{\min}, k) \tag{B.2}$$

$$= \prod_{i=1}^{N} a(x_{\min}, k) x_i^{-k}. \tag{B.3}$$

To solve for the normalizing constant a, I integrate the likelihood function over $x_{\min} \to \infty$ and set it equal to one assuming that $k > 1$:

$$1 = \int_{x_{\min}}^{\infty} a(x_{\min}, k) x^{-k} dx \tag{B.4}$$

$$1 = a(x_{\min}, k) \frac{1}{-k+1} x^{-k+1} \Big|_{x_{\min}}^{\infty} \tag{B.5}$$

$$1 = a(x_{\min}, k) \frac{1}{-k+1} \infty^{-k+1} - a(x_{\min}, k) \frac{1}{-k+1} x_{\min}^{-k+1} \tag{B.6}$$

$$\tag{B.7}$$

Since $\frac{1}{\infty} \to 0$, I can rearrange to solve for the normalizing constant:

$$1 = -a(x_{\min}, k) \frac{1}{-k+1} x^{-k+1} \tag{B.8}$$

$$-(k-1) = -a(x_{\min}, k) x_{\min}^{-k+1} \tag{B.9}$$

$$k - 1 = a(x_{\min}, k) \frac{1}{x_{\min}^{k-1}} \tag{B.10}$$

$$a(x_{\min}, k) = (k-1) x_{\min}^{k-1} \tag{B.11}$$

From here, one can substitute the normalizing constant into the original power law equation to solve for the MLE. But first, it is convenient to rearrange the power law equation to make the calculation easier:

$$f(x) = a(x_{\min}, k)x^{-k} \tag{B.12}$$

$$= (k-1)x_{\min}^{k-1}x^{-k} \tag{B.13}$$

$$= (k-1)x_{\min}^{k} - x_{\min}^{-}1x^{-k} \tag{B.14}$$

$$= \frac{(k-1)}{x_{\min}} \cdot \frac{x^{-k}}{x_{\min}^{-k}} \tag{B.15}$$

$$f(x) = \frac{(k-1)}{x_{\min}} \cdot \left(\frac{x}{x_{\min}}\right)^{-k} \tag{B.16}$$

$$\tag{B.17}$$

The likelihood L and the log-likelihood $\mathcal{L}$ are

$$L(x_i|x_{\min}, k) = \prod_{i=1}^{N} \frac{k-1}{x_{\min}} \cdot \left(\frac{x_i}{x_{\min}}\right)^{-k} \tag{B.18}$$

$$\mathcal{L}(x_i|x_{\min}, k) = \ln\left[\prod_{i=1}^{N} \frac{k-1}{x_{\min}} \cdot \left(\frac{x_i}{x_{\min}}\right)^{-k}\right] \tag{B.19}$$

$$= \sum_{i=1}^{N}\left[\ln(k-1) - \ln(x_{\min}) - k\ln\left(\frac{x_i}{x_{\min}}\right)\right] \tag{B.20}$$

$$\mathcal{L}(x_i|x_{\min}, k) = N\ln(k-1) - N\ln(x_{\min}) - k\sum_{i=1}^{N}\left[\ln\left(\frac{x_i}{x_{\min}}\right)\right]. \tag{B.21}$$

Differentiating the log-likelihood with respect to k and setting the derivative to zero

leads to the optimal value for k that maximizes the log-likelihood function:

$$\frac{d\mathcal{L}(x_i|x_{\min},k)}{dk} = 0 = \frac{N}{k-1} - \sum_{i=1}^{N}\left(\frac{x_i}{x_{\min}}\right) \tag{B.22}$$

$$\frac{N}{k-1} = \sum_{i=1}^{N}\left[\ln\left(\frac{x_i}{x_{\min}}\right)\right] \tag{B.23}$$

$$k-1 = N\left(\sum_{i=1}^{N}\left[\ln\left(\frac{x_i}{x_{\min}}\right)\right]\right)^{-1} \tag{B.24}$$

$$k = 1 + N\left(\sum_{i=1}^{N}\left[\ln\left(\frac{x_i}{x_{\min}}\right)\right]\right)^{-1} \tag{B.25}$$

B.1.1 Computing goodness-of-fit statistics and generating plots

This section describes additional details on fitting heavy tailed distributions to the steady state flux and rescaled EFM weight datasets. By defining generic power law methods, the log-likelihood and goodness-of-fit statistics for all distributions were computed using the `fitdistrplus`. Using the power law methods, I initialized an empty power law object and manually imputed the $x_{\min}$ and k parameters computed in the previous section. This allowed me to use `fitdistrplus` functions to numerically estimate the log-likelihood and the Kolmogorov-Smirnov (KS) statistic for all three distribution and generate the quantile-quantile (Q-Q), probability-probability (P-P), and complementary cumulative distribution function (CDF) plots.

B.2 Mapping genes to reactions

Table B.1: Gene-to-reaction mappings.

Reaction (#)	Gene	EntrezID	Enzyme
2	*ACER3*	55331	Alkaline ceramidase 3
3	*CERS1*	10715	Ceramide synthase 1
3	*CERS2*	29956	Ceramide synthase 2
3	*CERS3*	204219	Ceramide synthase 3
3	*CERS4*	79603	Ceramide synthase 4
3	*CERS5*	91012	Ceramide synthase 5
3	*CERS5*	253782	Ceramide synthase 6
4	*SPHK2*	56848	Sphingosine kinase 2
5	*SGPP1*	81537	Sphingosine-1-phosphate phosphatase 1
5	*SGPP2*	130367	Sphingosine-1-phosphate phosphatase 2
8	*SGMS1*	259230	Sphingomyelin synthase 1
8	*SGMS2*	166929	Sphingomyelin synthase 2
9	*SMPD2*	6610	Sphingomyelin phosphodiesterase 2
9	*SMPD3*	55512	Sphingomyelin phosphodiesterase 3
9	*SMPD4*	55627	Sphingomyelin phosphodiesterase 4
10	*ASAH2*	56624	N-acylsphingosine amidohydrolase 2
11	*SPHK2*	56848	Sphingosine kinase 2
12	*PLPP1*	8611	phospholipid phosphatase 1
17	*SGPP1*	81537	Sphingosine-1-phosphate phosphatase 1
18	*SPHK2*	56848	Sphingosine kinase 2
19	*CERS1*	10715	Ceramide synthase 1
19	*CERS2*	29956	Ceramide synthase 2
19	*CERS3*	204219	Ceramide synthase 3
19	*CERS4*	79603	Ceramide synthase 4
19	*CERS5*	91012	Ceramide synthase 5
19	*CERS5*	253782	Ceramide synthase 6
20	*ASAH1*	427	N-acylsphingosine amidohydrolase 1
25	*SMPD1*	6609	Sphingomyelin phosphodiesterase 1
26	*ASAH2*	56624	N-acylsphingosine amidohydrolase 2
27	*SPHK1*	8877	Sphingosine kinase 1
27	*SPHK2*	56848	Sphingosine kinase 2
28	*SMPD2*	6610	Sphingomyelin phosphodiesterase 2
29	*ASAH2*	56624	N-acylsphingosine amidohydrolase 2
30	*SPHK1*	8877	Sphingosine kinase 1
36	*SPHK2*	56848	Sphingosine kinase 2
40	*UGCG*	7357	UDP-glucose ceramide glucosyltransferase
43	*B4GALT6*	9331	Beta-1,4-galactosyltransferase 6
47	*GBA1*	2629	Glucosylceramidase beta 1

Table B.1 continued.

Reaction (#)	Gene	EntrezID	Enzyme
49	*SMPD1*	6609	Sphingomyelin phosphodiesterase 1
50	*ASAH1*	427	N-acylsphingosine amidohydrolase 1
53	*SPHK1*	8877	Sphingosine kinase 1
54	*PLPP1*	8611	phospholipid phosphatase 1
54	*PLPP2*	8612	phospholipid phosphatase 2
55	*SGMS1*	259230	Sphingomyelin synthase 1
57	*SMPD2*	6610	Sphingomyelin phosphodiesterase 2
58	*ACER2*	340485	Alkaline ceramidase 2
58	*ACER3*	55331	Alkaline ceramidase 3
59	*CERK*	64781	Ceramide kinase
60	*PLPP1*	8611	phospholipid phosphatase 1
60	*PLPP2*	8612	phospholipid phosphatase 2
62	*PLPP1*	8611	phospholipid phosphatase 1
62	*PLPP2*	8612	phospholipid phosphatase 2
62	*PLPP3*	8613	phospholipid phosphatase 3
63	*CERK*	64781	Ceramide kinase
69	*PLPP1*	8611	phospholipid phosphatase 1
69	*PLPP2*	8612	phospholipid phosphatase 2
69	*PLPP3*	8613	phospholipid phosphatase 3

B.3 Fitted distribution from the poweRlaw package

The following data were fitted to heavy- and light-tailed distributions using the method of Clauset et al.[223] as implemented in the `poweRlaw` package:[222]

1. All steady state fluxes within a given sample condition.
2. All rescaled EFM weights within a given sample condition.
3. The unperturbed $V_{\max}$ parameters within the Alzheimer's disease and control condition.

Goodness-of-fit was assessed by bootstrapping 1000 synthetic samples for each empirical distribution and computing a P-value defined as the fraction of synthetic KS statistics greater than that from the empirical data. Data plausibility following a given distribution was defined at a $P > 0.10$ cutoff. Note that the lower bound for each heavy-tailed distribution was computed by finding the smallest value of $x_{\min}$ that minimized the KS statistic.

B.3.1 All fluxes within given condition sample

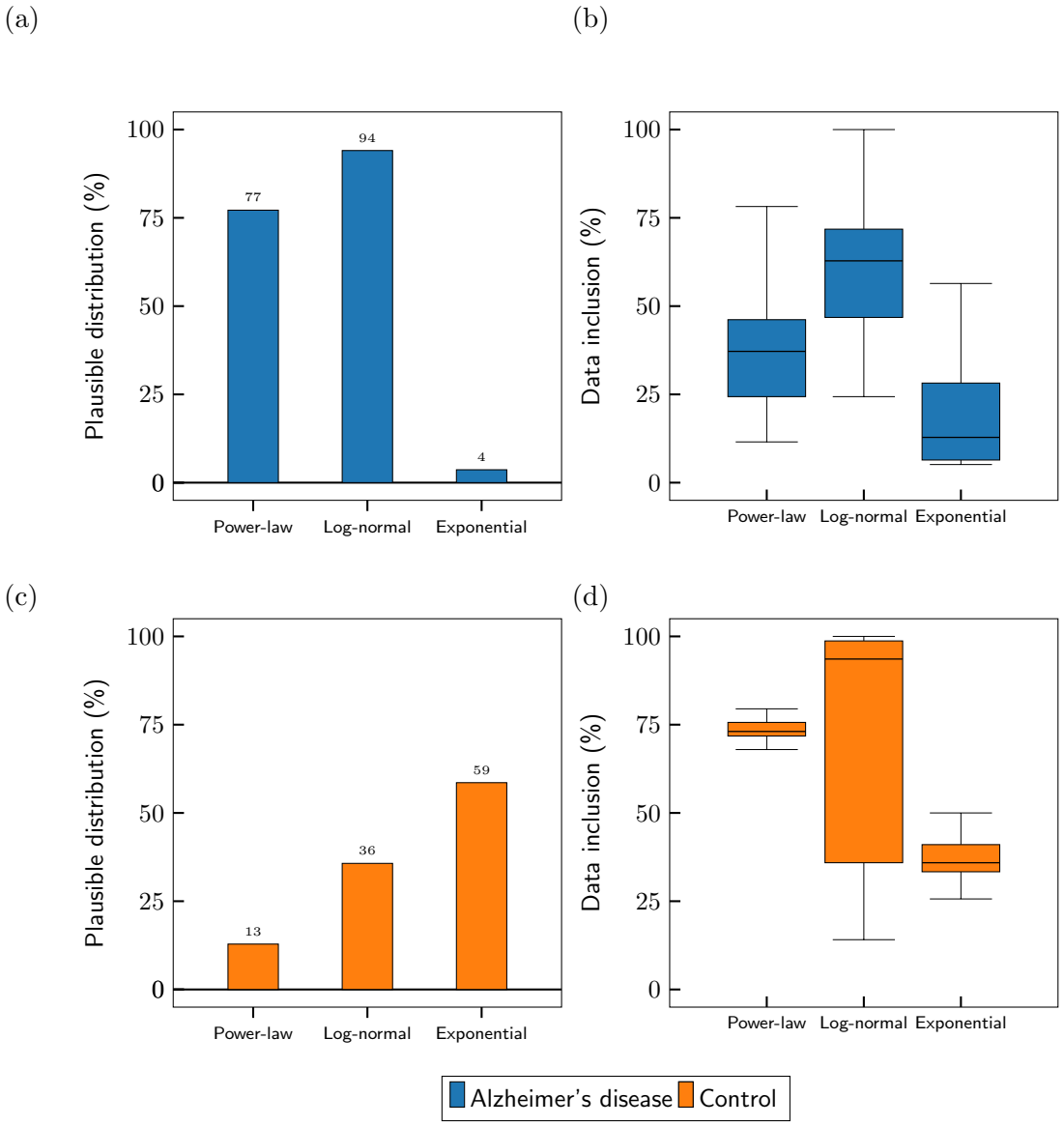


Figure B.1: Fitted distributions of all fluxes within each condition sample using Method two. (a-b) Percentage of samples plausibly belonging to the given distribution and percentage of included data for each fit for the Alzheimer's disease samples. (c-d) Percentage of samples plausibly belonging to the given distribution and percentage of included data for each fit for the control samples.

B.3.2 All rescaled EFM weights within given condition sample

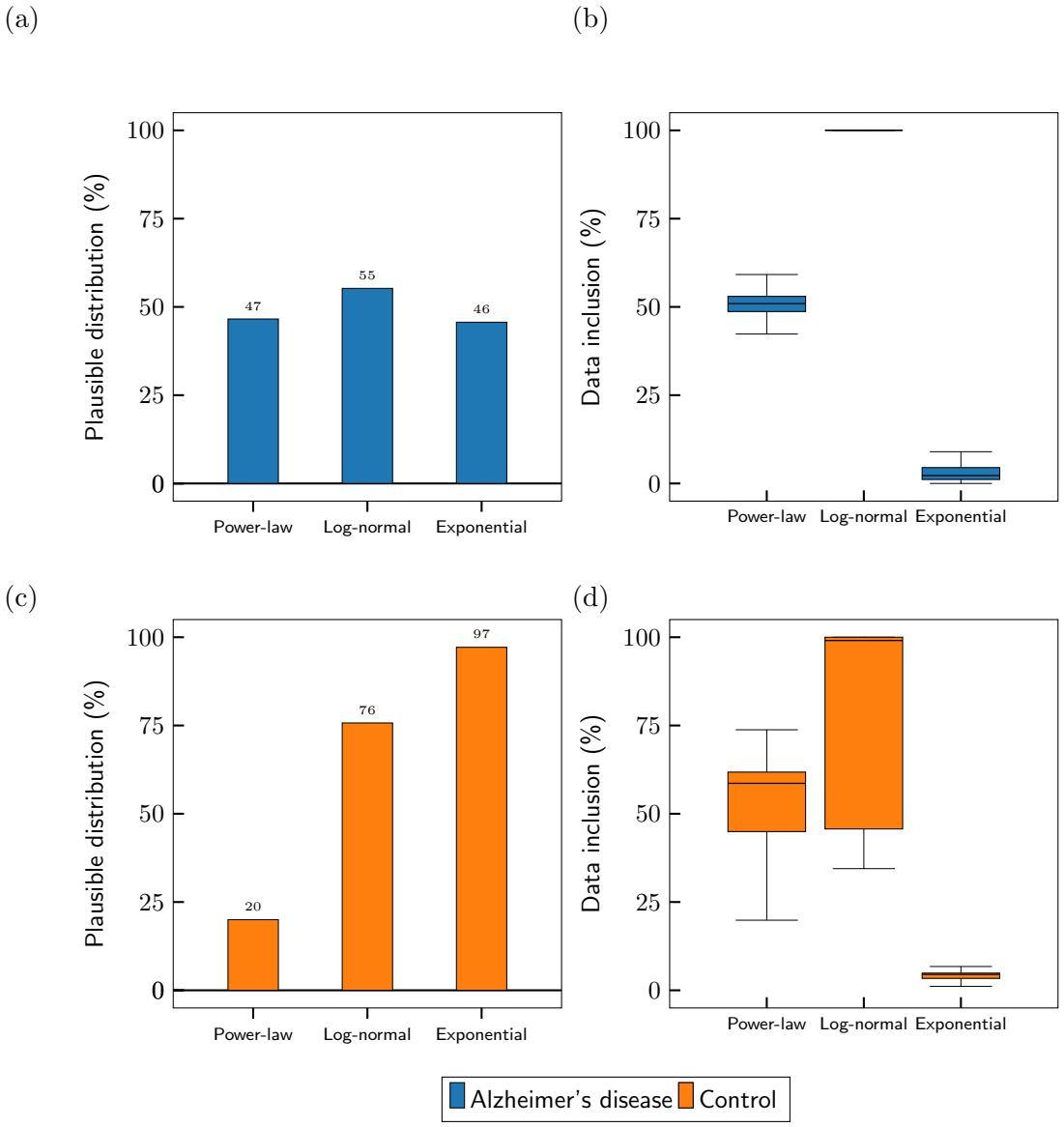


Figure B.2: Fitted distributions of rescaled EFM weights for each condition sample using Method two. (a-b) Percentage of samples plausibly belonging to the given distribution and percentage of included data for each fit for the Alzheimer's disease samples. (c-d) Percentage of samples plausibly belonging to the given distribution and percentage of included data for each fit for the control samples.

B.3.3 Distribution fitting for $V_{\max}$ parameters in the unperturbed models

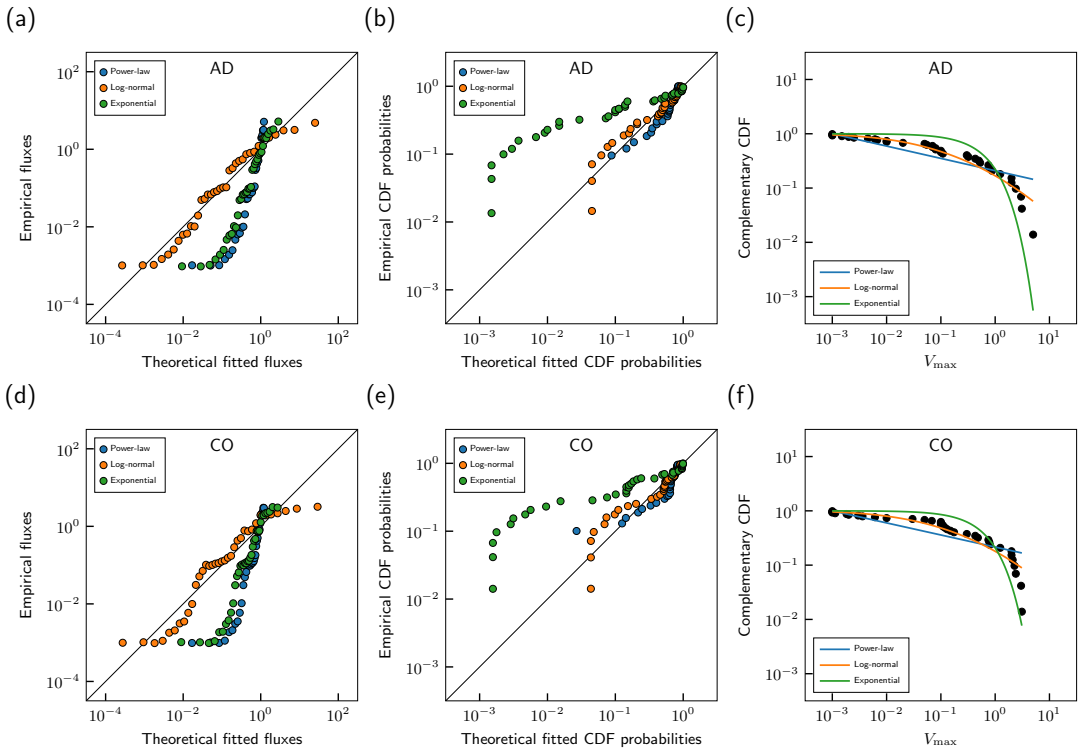


Figure B.3: Fitted distributions to the $V_{\max}$ parameters within unperturbed conditions using Method one. (a-c) Q-Q plot, P-P plot, and complementary CDF plot of Alzheimer's disease samples. (d-f) Q-Q plot, P-P plot, and complementary CDF plot of control samples.

C

Atomic elementary flux modes explain the steady state flow of metabolites in large-scale flux networks

C.1 Details associated with computing the (A)CHMC stationary distribution

Following basic Markov chain theory, the stationary distribution of the CHMC, π is proportional to the eigenvector of the CHMC transition probability matrix. The eigenvector can be solved by direct or iterative numerical methods. The choice of algorithm becomes increasingly important as the size of the CHMC transition matrix grows exponentially in highly-connected, large-scale networks. I note that iterative eigenvector solving algorithms from some packages (e.g. Arnoldi method from Arnoldi.jl, GMRES from IterativeSolvers.jl) resulted in negative eigenvector coefficients or NaN values. These are possibly numerical errors and occurred more frequently in larger CHMC matrices. Repeatedly re-running these methods sometimes resulted in a correct eigenvector, albeit with significantly longer computation times. I observed that the linear solver from the Julia package LinearSolve.jl efficiently solved for the correct eigenvector without error using the default parameters. Hence, I use this package in my implementation and recommend using it to compute the stationary distribution of large CHMCs and ACHMCs.

D

Atomic elementary flux mode analysis of five large-scale metabolic networks

D.1 Datasets

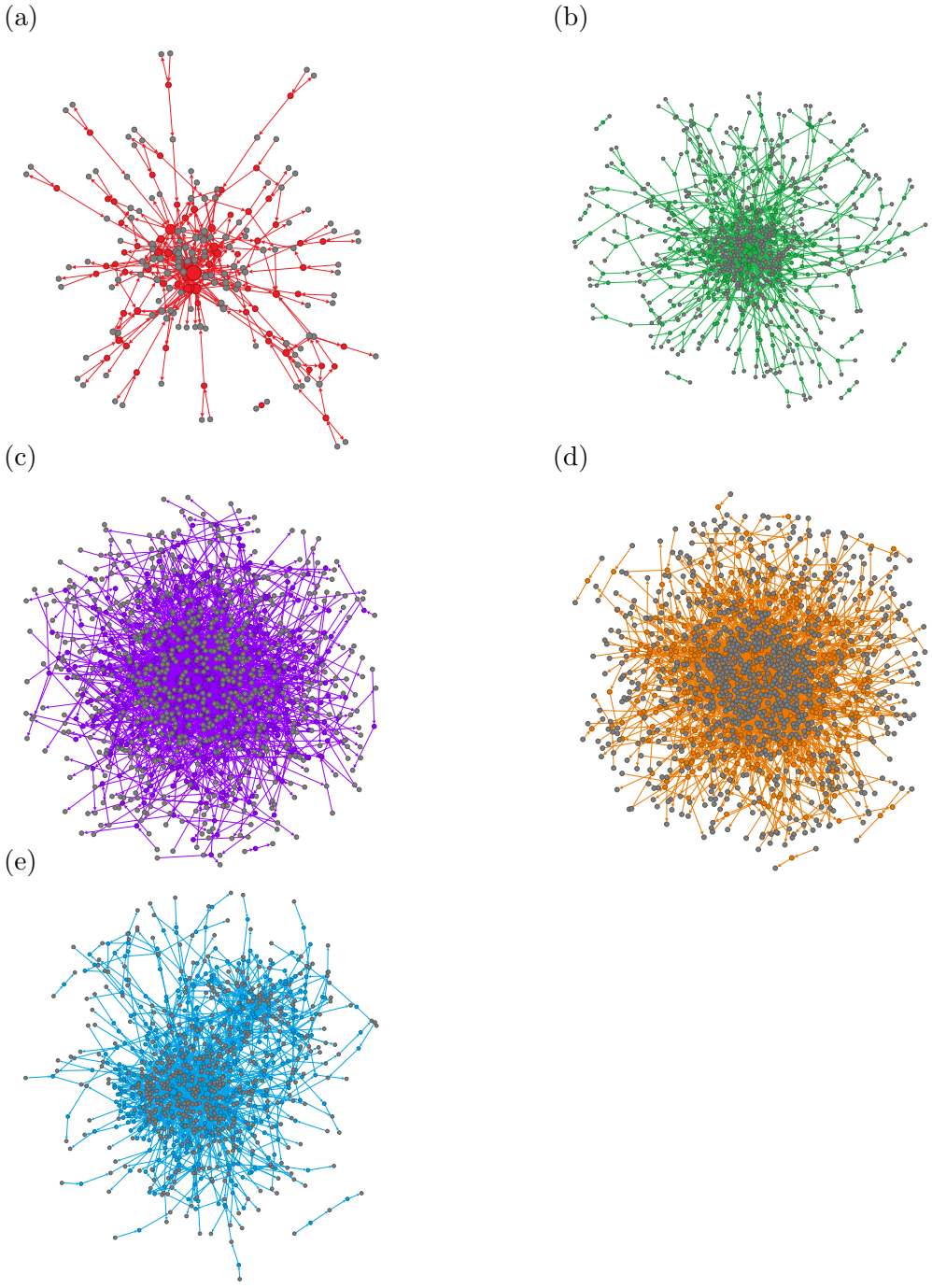


Figure D.1: Network visualization of the five GEMs after pre-processing. Metabolites (non-grey nodes) are connected to reaction entities (grey nodes) with node sizes proportional to their degree. Graphs were constructed in Gephi version 0.10 using the Yifan-Hu proportional network layout with default parameters. (a) E. coli core. (b) iAB RBC 283. (c) iIT341. (d) iSB619. (e) HepG2.

D.2 Pre-processing details

Table D.1: Pre-processing details for the five GEMs.

	E. coli core	iAB RBC 283	iIT341	iSB619	HepG2
Reactions with non-integer stoichiometries	2	1	3	34	8
Pseudometabolites with no known structures	1	68	52	108	33
Reactions with pseudometabolites	2	142	93	204	46
Reactions exceeding RXNMapper token limit	0	1	4	14	3
Transaminase reactions (manually corrected)	0	1	8	11	10

D.3 Results

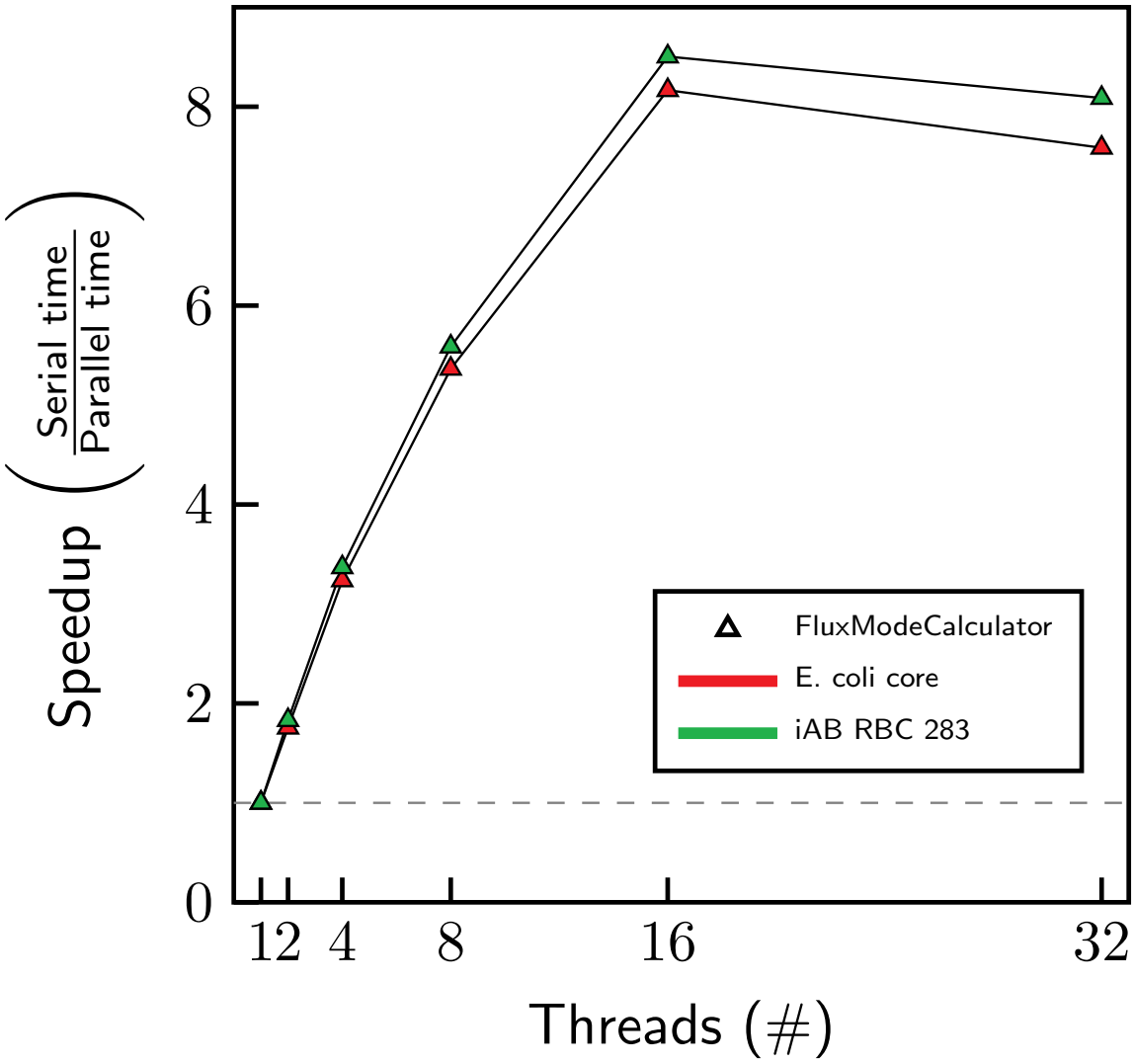


Figure D.2: FluxModeCalculator scaling across multiple threads.

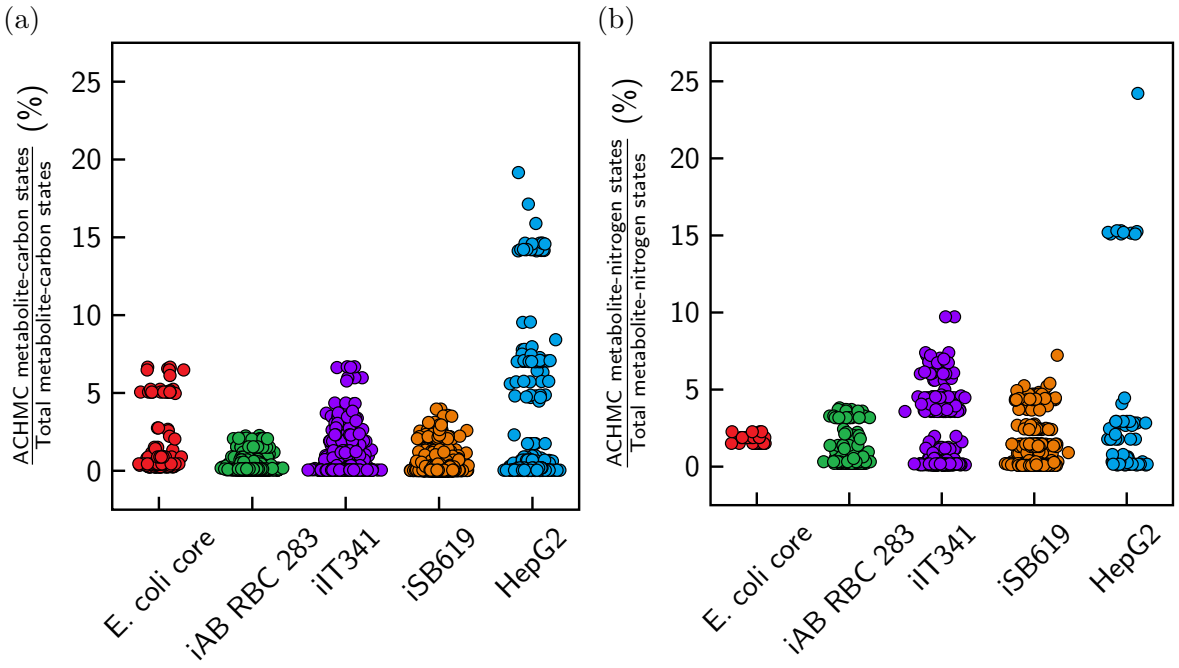


Figure D.3: Most ACHMCs traverse less than 10% of the total metabolite-carbon/nitrogen state space. Each point corresponds to an ACHMC rooted on a given source metabolite carbon (a) or nitrogen (b).

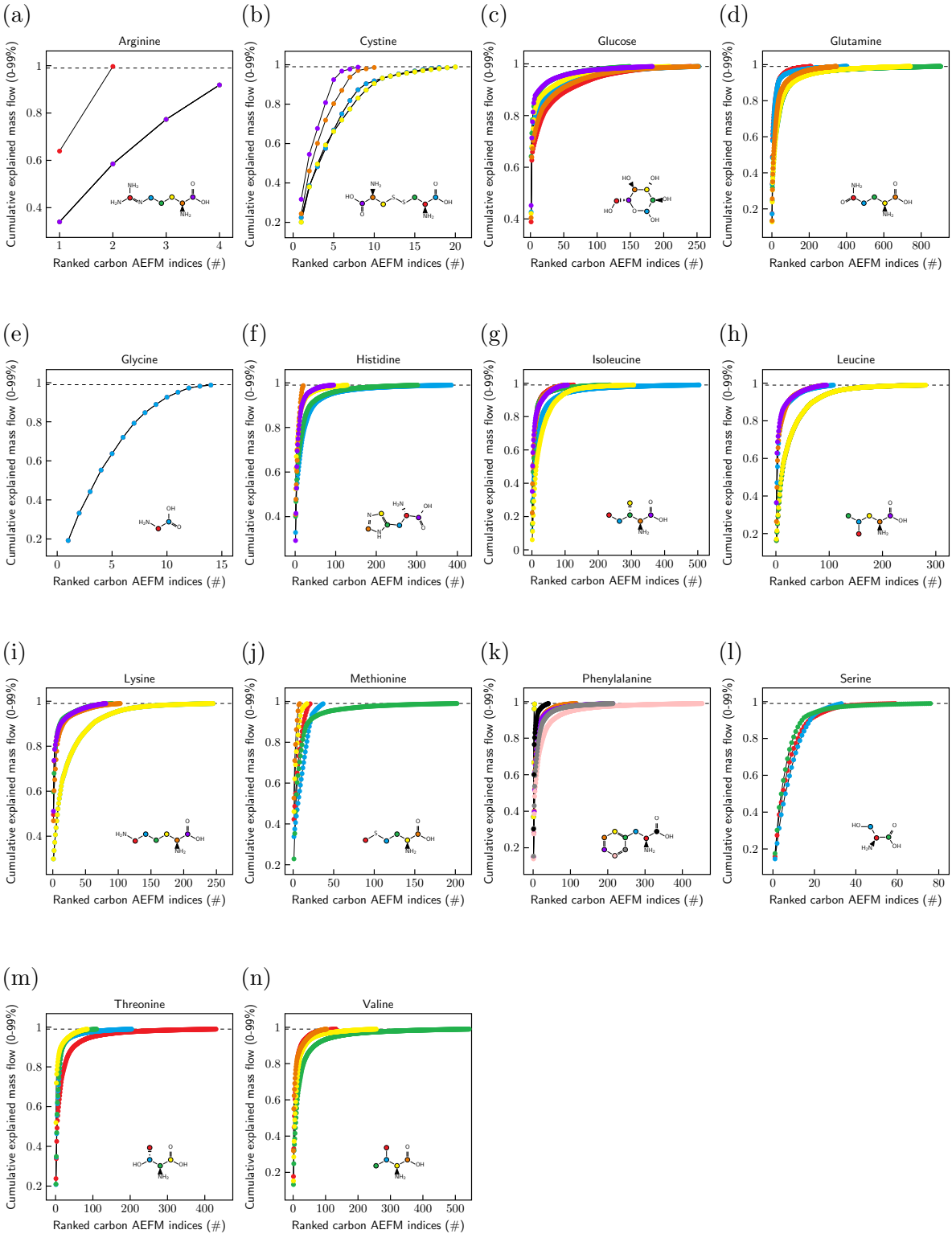


Figure D.4: Cumulative explained carbon mass flow of input glucose and amino acid carbons from the HepG2 model. The cumulative explained mass flow is truncated at 99% to avoid displaying AEFMs that carry little atomic mass flow.

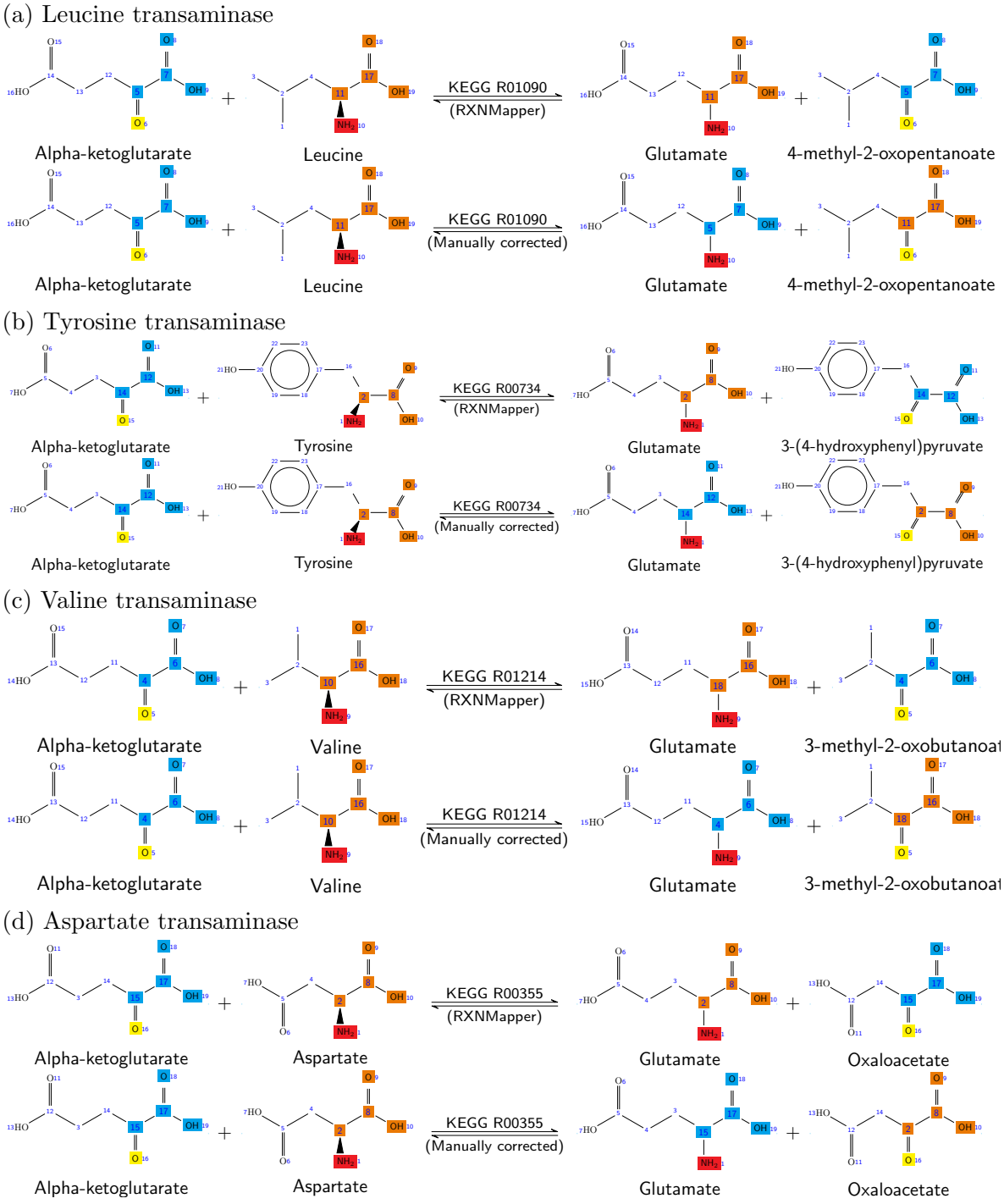


Figure D.5: Examples of incorrectly atom-mapped transaminase reactions present in one or more of the five networks. (a) Leucine transaminase. (b) Tyrosine transaminase. (c) Valine transaminase. (d) Aspartate transaminase.

Curriculum Vitae

Justin G. Chitpin was born in Toronto and grew up in Ottawa, Canada. He started his undergraduate studies at the University of Ottawa studying biomedical science with specialization in biostatistics. Upon completing his second year, he switched to the newly inaugurated Translational and Molecular Medicine program. Here, he discovered high-throughput sequencing techniques and the bioinformatics methods used to analyze large-scale data. Under the supervision of Dr. Theodore J. Perkins, he completed his Honour's project titled "RECAP reveals the true statistical significance of ChIP-seq peaks." Interested in learning about other types of high-throughput data, he remained at the University of Ottawa under Dr. Perkins to complete a two-year Master's degree in biochemistry with specialization in bioinformatics. He developed methods to analyze lipidomics data from targeted mass spectrometry data acquisition modes with his Masters thesis on "Bioinformatics for quantitative, targeted lipidomics discovery." Following his desire to transition to systems biology modelling, he decided to remain with Dr. Perkins at the University of Ottawa to pursue a PhD in the newly created biochemistry program with specialization in bioinformatics. His PhD focused on developing new methods to analyze metabolic flux data with his thesis on "Advancing elementary flux mode analysis for analyzing large-scale metabolic flux networks under steady state." Currently, Justin is a postdoctoral research fellow at the University of Queensland's Australian Institute for Bioengineering and Nanotechnology in Brisbane, Australia.