In silico prediction of metabolic profiles for improved biomarker discovery and metabolic phenotyping

Juliette Cooke

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

Institut National Polytechnique de Toulouse (Toulouse INP)

Discipline ou spécialité :

Infectiologie, Physipoathologie, Toxicologie, Génétique et Nutrition

---

Présentée et soutenue par :

Mme JULIETTE COOKE

le lundi 18 décembre 2023

Titre :

Prédiction in silico de profils métaboliques pour améliorer la découverte des biomarqueurs et le phénotypage métabolique

---

Ecole doctorale :

Sciences Ecologiques, Vétérinaires, Agronomiques et Bioingénieries (SEVAB)

Unité de recherche :

Toxicologie Alimentaire (ToxAlim)

Directeur(s) de Thèse :

M. FABIEN JOURDAN

Rapporteurs :

M. KARL BURGESS, UNIVERSITY OF EDINBURGH

MME AUDREY LE GOUELLEC, UNIVERSITE GRENOBLE ALPES

MME SOPHIE COLOMBIE, INRA VILLENAVE D'ORNON

Membre(s) du jury :

M. JEAN-CHARLES PORTAIS, INSA TOULOUSE, Président

M. FABIEN JOURDAN, INRA TOULOUSE, Membre

M. LAURENT LE CAM, INSERM MONTPELLIER, Membre

Summary

Human metabolic phenotyping can detect abnormal physiological changes via metabolites circulating in biofluids (plasma, urine). These metabolic profiles can be obtained by classical clinical assays (small targeted number of molecules) or by a more global approach aimed at measuring a wide range of endogenous molecules without a priori knowledge, known as metabolomics.

However, metabolomics approaches cannot cover the entire human metabolome with a single analytical technique. It is therefore essential to plan and optimise metabolomics experiments to ensure that the covered metabolome will be as relevant as possible for the condition studied. The hypothesis of this thesis work is that global modelling of metabolism makes it possible to simulate a metabolic disturbance by being free from the constraints of coverage and observability of metabolomics, and thus assist the experimental design involving these techniques.

The other challenge that metabolomics applied to biofluids faces is understanding how to link metabolic profiles with the molecular metabolic perturbations that caused them. In this context, the thesis work aims at proposing a modelling method to simulate, from a molecular event (e.g. inhibition of an enzymatic activity), the metabolic profile signalling the physiological drift.

The central objective of the thesis is therefore to create a predictive system which can simulate metabolic perturbations, and to recommend the most changed metabolites associated with them. For this, the project consists in modelling human metabolism by simulating the exchanges involving all the metabolic reactions that can take place in humans. This modelling, known as constraint-based modelling, makes it possible to simulate metabolic fluxes (rate of production and consumption of metabolites) and thus to predict which metabolites will be present or not in biofluids.

In this thesis, a constraint-based modelling approach is developed and applied to predict *in silico* profiles of metabolites that are more likely to be

differentially abundant under a given metabolic perturbation (e.g. due to a genetic disease) using flux simulation. In genome-scale metabolic networks (large networks containing metabolic, enzymatic and genetic data and how they are linked together), the fluxes through reactions which transport metabolites externally (called exchange reactions) can be simulated and compared between control and disease conditions in order to calculate changes in metabolite import and export. These import/export flux differences are expected to induce changes in circulating biofluid levels of those metabolites, which can then be interpreted as potential metabolites of interest. SAMBA (SAMpling Biomarker Analysis), developed for this project, is an approach which simulates fluxes in exchange reactions following a metabolic perturbation using random sampling, compares the simulated flux distributions between the baseline and modulated conditions, and ranks predicted differentially exchanged metabolites as potential biomarkers for the perturbation.

The project's results show that there is a good fit between simulated metabolic exchange profiles and experimental differential metabolites detected in plasma, such as patient data from the disease database OMIM (Online Mendelian Inheritance in Man), and metabolic trait-SNP (Single Nucleotide Polymorphism) associations found in mGWAS (metabolite genome-wide association study) studies. These metabolic profile recommendations can provide insight into the underlying mechanism or metabolic pathway perturbation lying behind observed metabolite differential abundances, and suggest new metabolites as potential avenues for further experimental analyses.

Résumé

Le phénotypage métabolique humain permet de déceler des dérives physiologiques anormales via des molécules circulantes dans les biofluides (plasma, urines). Ces profils métaboliques peuvent être obtenus par des dosages cliniques classiques (petit nombre ciblé de molécules) ou par une approche plus globale visant à mesurer sans a priori une gamme large de molécules endogènes : la métabolomique.

La métabolomique ne permet pas de couvrir avec une seule technique analytique l'ensemble du métabolome humain. Il est donc indispensable de planifier et d'optimiser les expériences de métabolomique pour s'assurer que le métabolome sera couvert de la façon la plus pertinente possible pour chaque condition. L'hypothèse de ce travail de thèse est que la modélisation globale du métabolisme permettrait de simuler une perturbation métabolique autorisant de dépasser les contraintes actuelles de couverture et d'observabilité de la métabolomique, et ainsi permettre d'assister le design expérimental impliquant ces techniques.

L'autre défi auquel fait face la métabolomique appliquée aux biofluides est d'établir un lien entre les profils métaboliques et les perturbations moléculaires métaboliques qui en sont à l'origine. Dans ce contexte, ce travail de thèse vise à proposer une méthode de modélisation permettant de simuler, à partir d'événements moléculaires ou biochimiques, le profil métabolique diagnostique d'une dérive physiologique.

L'objectif central de la thèse est donc de créer une approche prédictive qui permet de simuler des perturbations métaboliques et d'y associer les métabolites qui sont le plus affectés. Pour cela, le projet consiste à modéliser le métabolisme humain en modélisant les échanges impliquant l'ensemble des réactions métaboliques qui peuvent avoir lieu chez l'homme. Cette modélisation, dite sous contrainte, permet de simuler les flux métaboliques (taux de production et de consommation de métabolites) pour prédire quels métabolites se retrouvent ou

non dans les biofluides.

Dans cette thèse, une approche de modélisation sous contraintes est développée et appliquée pour prédire *in silico* les profils métaboliques qui sont les plus susceptibles d'être différentiellement abondants lors d'une perturbation métabolique donnée en utilisant la simulation de flux. Dans les réseaux métaboliques à l'échelle du génome (réseaux contenant des données génomiques, enzymatiques, et métaboliques ainsi que les liens qui les relient), les flux des réactions d'échange, également connues sous le nom de réactions qui transportent les métabolites vers l'extérieur, peuvent être simulés et comparés entre les conditions de contrôle et de maladie afin de calculer les changements dans l'import et l'export des métabolites. Ces différences de flux d'import et d'export devraient induire des changements dans le niveau de ces métabolites dans les biofluides circulants, qui peuvent alors être interprétés comme des métabolites d'intérêt potentiels. SAMBA (SAMpling Biomarker Analysis), développé pour ce projet, est une approche qui simule les flux dans les réactions d'échange suite à une perturbation métabolique en utilisant le random sampling (ou échantillonnage aléatoire), compare les distributions de flux simulées entre la condition de base et la condition modulée, et ordonne les métabolites prédits comme différentiellement échangés en tant que biomarqueurs potentiels de la perturbation.

Les résultats du projet montrent qu'il existe une bonne correspondance entre les profils d'échanges métaboliques simulés et les métabolites différentiels expérimentaux détectés dans le plasma, tels que les données sur les patients de la base de données sur les maladies OMIM (Online Mendelian Inheritance in Man), et les associations entre trait métabolique et SNP (Single nucleotide polymorphism) trouvées dans les études mGWAS (metabolite genome-wide association study). Ces recommandations de profils métaboliques peuvent donner un aperçu du mécanisme métabolique ou de la perturbation de la voie métabolique qui se cache derrière les différences de métabolites observées, et

suggérer de nouveaux métabolites comme pistes potentielles pour des analyses expérimentales plus poussées.

Acknowledgements

First and foremost I would like to thank my three thesis rapporteurs, Karl Burgess, Audrey Le Gouëllec, and Sophie Colombié, for taking the time to read my manuscript and providing their valuable feedback. I also thank Jean-Charles Portais and Laurent Le Cam for accepting to be examiners in my thesis jury. I am also thankful to Sabine Pérès and Timothy Ebbels for being a part of my PhD committee, guiding and helping me with my project.

I'm extremely grateful to Fabien Jourdan for supervising and guiding me throughout this 3-year journey, for his patience, and for his availability and flexibility despite important life events (for both of us!). Without his help and guidance, I would have been as lost as he is whenever I make a reference that is too nerdy.

Many thanks to everyone who is a part of the collaboration with Imperial College London, Cecilia Wieder, Timothy Ebbels, Jake Bundy, and Rachel Lai, for their amazing work as well as helpful insight. Your fresh perspectives are something that really help push our collaborative work further, as well as my PhD project.

I would like to express my deepest appreciation to the MetExplore team for all the wonderful scientific and fun moments we have shared together: Fabien (again), Nathalie, Clément, Florence, Ludo, Louison, Jean-Clément, Maximilian, Marion, Elva, Bénédict, Maxime and Pablo. Fabien, Nathalie, and Clément, thank you for your endless support in proofreading and providing feedback, in sharing your knowledge, and in being scientific role models. Florence, Jean-Clément, Marion, and Ludo, thank you for your technical expertise, organisation, and dedication to the wonderful tool that is MetExplore. Louison, Maximilian, Elva, and Bénédict, thank you for always teaching me new things through your code and presentations. Pablo, thank you for pointing me in the right direction and for all of your interesting discussions. Maxime, thank you for being an amazing office neighbour, as well as for our long discussions about the endless possibilities

of science. To all of you, thank you for creating a wonderful atmosphere for learning, work, discussions and fun, despite COVID changing how we worked. Our discussions on video games, board games, films and shows are something I really cherish. And thank you for showing me the way of beer and coffee (with moderation)!

I would also like to thank the MeX team in its entirety for not only the scientific discussions but also the celebrations and gatherings, at the lab (pétanque!) and at the Caporal bar. Thank you Elodie for being an awesome office neighbour during my third year! I'd also like to acknowledge everyone at the ToxAlim lab, with special thanks to everyone in the Graines de ToxAlim group for bringing us closer through board games and beer.

Thank you to all those involved with the doctoral school SEVAB and the Ministry funding, without whom this thesis would not have been possible financially and administratively. I'd also like to thank my classmates as well as the members of the teaching staff at Université Paul Sabatier during my Bioinformatics Master.

I would also like to thank my parents, Gabi, the rest of my family (both old and new), and my friends, especially Nathalie, Marion, and Marie-Alphée, for supporting me and being patient throughout these past years. A huge thank you to them (and Echo, also known as Echo-chan, Boubou, la Boubz) for their initiative and proactivity in getting me to go outside (which is no easy feat)! While some would also thank the lab canteen, I would like to thank my kitchen robot for helping me avoid food-related depression during these three years.

Finally, a special thank you to Matthieu for not only supporting me emotionally but also helping me scientifically. I especially thank him for his patience with my R scripts and my repeated "So I have this... and I want to get this..." questions! It is wonderful to be able to share that with someone so close, and doing our PhDs at the same time would have been impossible without his patience and positivity.

Quotes

I thought there couldn't be anything as complicated as the universe, until I started reading about the cell.

Systems Biologist Eric de Silva

-astrophysicist by training-

Imperial College London

In theory, theory and practice are the same.
In practice, they are not.

Albert Einstein

Art is not one great act of creation, but many small ones. When you read one of my poems, you fail to see the weeks of careful work it took me to build it - the thinking, the scratched-out words, the pages I burned in disgust. All you see, in the end, is what I want you to see.

Samantha Shannon

The Priory of the Orange Tree

Many fall in the face of chaos... but not this one, not today.

Darkest Dungeon

Red Hook Studios

Note

In PDF form, all citations, contents titles, figure and table labels, and acronyms are hyperlinked and clickable for ease of reading. Citations in the bibliography section contain backlinks to return to the cited location(s).

Contents

List of Figures

List of Tables

List of Acronyms

ACADS Acyl-CoA Dehydrogenase Short chain . . . 110
ACF autocorrelation function . . . 135
ACHR Artificial Centering Hit-and-Run . . . 82
ATP Adenosine TriPhosphate . . . 24
CBM constraint-based modelling . . . 54
ChEBI Chemical Entities of Biological Interest . . . 43
EC Enzyme Commission . . . 43
FN false negative . . . 153
FP false positive . . . 153
FBA Flux Balance Analysis . . . 60
FVA Flux Variability Analysis . . . 62
GO Gene Ontology . . . 142
GEM genome-scale metabolic model . . . 45
GPR Gene Protein Reaction . . . 46
GSEA gene set enrichment analysis . . . 123
GSMN genome-scale metabolic network . . . 45
GWAS genome-wide association study
HMDB Human Metabolome DataBase . . . 43
IEM Inborn Errors of Metabolism . . . 25
KEGG Kyoto Encyclopedia of Genes and Genomes . . . 42
KO knock-out . . . 76
KD knock-down . . . 78
LC liquid chromatography
mGWAS metabolite genome-wide association study . . . 28
MS mass spectrometry . . . 33
MSEA metabolite set enrichment analysis . . . 123
MSUD maple syrup urine disease . . . 26
MUT mutant . . . 77

NADH Nicotinamide Adenine Dinucleotide + Hydrogen

NADPH Nicotinamide Adenine Dinucleotide Phosphate + Hydrogen

NES normalised enrichment score 146

NMR nuclear magnetic resonance 33

ODE ordinary differential equations 53

OMIM Online Mendelian Inheritance in Man 25

ORA overrepresentation analysis 123

PCA principal component analysis 36

PLS-DA partial least squares discriminant analysis 36

PSRF potential scale reduction factor 135

SAMBA SAMpling Biomarker Analysis 91

SCD Stearoyl-CoA 9-desaturase 110

SNP single nucleotide polymorphism 27

SBML Systems Biology Markup Language 44

TN true negative 153

TP true positive 153

WT wild type 75

Chapter I

Introduction: Measuring and modelling metabolism

1 Metabolomics and its uses in human health

1.1 Metabolism

Metabolism is a crucial biological process which ensures that cells have the energy and components required to survive, function, and grow. It consists of the set of chemical reactions which break down molecules into smaller compounds known as metabolites (any molecule with a molecular weight less than 1.5 kDa), produce energy, build and repair tissue, eliminate metabolic waste, and provide the ability to respond to the surrounding environment. These metabolites are involved in every living organism's metabolism and are produced, degraded, and transformed via biochemical reactions, the majority of which take place inside cells. Different cellular compartments such as mitochondria, cytoplasm, and nucleus all have specific enzymatic expression profiles, which can lead to metabolites only being present within certain compartments. Some metabolites are not in contact with each other as they are not used in every compartment and only traverse these areas through transport mechanisms (passive or active).

Metabolism is globally grouped into two processes: catabolism and

anabolism. While catabolism is the process of breaking down compounds to release energy, anabolism consumes this energy to synthesise larger compounds from smaller molecules. Breaking down glucose in cellular respiration to produce energy in the form of Adenosine TriPhosphate (ATP) is an example of a catabolic process. This mechanism is the essential first step to being able to harness external molecules and transform them into something usable by the organism. Conversely, the synthesis of proteins from amino acids is an example of an anabolic process which consumes energy. Anabolism can be viewed as the second step in providing the organism with the ability to survive and grow. It uses the energy released from catabolic reactions and building blocks to biosynthesise new larger molecules, which can be functional or for storage purposes. Anabolism is highly controlled and regulated by the cell which helps avoid wasting energy in infinite loops of synthesis and degradation.

Being able to observe and measure the presence, quantities and variations of these small molecules is of major importance in human health. First, and most evidently, any disease related directly to metabolic disruptions can be efficiently identified and described by measuring the concentrations of specific metabolites, or biomarkers, in the blood or urine. A second, more recently developed way of viewing a patient's metabolism is by profiling a large range of metabolites at once, to gain a global perspective of the current state of metabolic activity.

1.2 Biomarkers

Biomarkers are measurable indicators of a given biological state, and can be anything from externally measurable markers such as temperature or weight, to internally evaluated markers in biofluids like blood or urine. When applied to human health research, a subset of the metabolome can be considered as metabolic biomarkers of a given pathology if this subset is statistically shared by a homogeneous group of patients in comparison to control subjects or another

group of patients not affected by the pathology under study. A well-known example is the level of glucose in diabetic patients compared with non-diabetic individuals. In non-diabetic people, the average fasting level of blood sugar is below 100 mg/dL, whereas a higher fasting blood glucose concentration indicates either a prediabetic or diabetic state. For type I diabetes, the pancreas cannot produce insulin, which is the activator for the glucose channel to open. For type II diabetes, the pancreas produces insulin but cells have become resistant to its effect. In short, for both types of diabetes, this results in an accumulation of glucose in the bloodstream. Because glucose is easily measurable using many different types of glucose monitors, it is a useful biomarker of a diabetic state, as well as an indicator that can be tracked over time once diagnosed.

In clinical settings, biomarkers are traditionally detected using targeted bioassays, which result in measurements for a small number of well-characterised diagnostic metabolites. Specific sets of biomarkers are known to be associated with certain diseases, meaning that future cases of these diseases can be easily diagnosed using previously observed information, and they are useful for monitoring disease progression as well as predicting the onset of degenerative diseases. Biomarker discovery is a large part of human health research and pharmaceutical studies, for use as intermediate diagnostic markers and potential drug targets. A major advantage of using biomarkers in disease diagnosis is the easy accessibility of these biomarkers in biofluids, as opposed to more invasive approaches like organ biopsies.

These metabolic biomarkers are especially useful for diagnosing a subset of diseases called Inborn Errors of Metabolisms (IEMs). IEMs are rare genetic mutations affecting enzyme-coding genes. They tend to affect systems such as carbohydrate metabolism, the urea cycle, amino acids, and mitochondrial functions, and generally result in clinically significant symptoms. Online databases such as Online Mendelian Inheritance in Man (OMIM) [1] contain associations between various diseases and metabolic biomarkers, using past

patient case reports and publications. However, for diseases such as IEMs, the data is based on very few patients and can have a large inter-patient variability due to variation in parameters such as the weight of patients. Treatments are generally swiftly tailored to the specific disorder once a diagnosis is made due to the often young age of the patients, with the goal of eliminating the build-up of excess or toxic metabolites that result from this metabolic dysregulation. Due to the direct link of these diseases with metabolism, biomarkers have been identified over the years in patients with each disease. These biomarkers are usually specific to a given disorder and are therefore ideal candidates for use as diagnostic indicators for IEMs. For example, maple syrup urine disease (MSUD), named after the odour it gives to the urine of patients with this disease, is an IEM caused by a genetic mutation in enzymes involved in the catalysis of branched-chain amino acids. The major metabolic biomarkers for MSUD are elevated levels of leucine, isoleucine, and valine, which can be measured in newborn serum for an early diagnosis [2].

1.3 Metabolic profiles

In contrast to biomarker-level measurements of metabolites, entire lists of metabolites can be measured, identified and quantified (most often using relative quantification), constituting a metabolic profile representative of a given state. This is of course thanks to newer experimental approaches which can measure not only an entire class of metabolites but also multiple classes at once. Metabolic profiling consists of the measurement of these profiles to evaluate the response to physiological, pathophysiological or otherwise environmental stimuli, as well as measure the behaviour of metabolism in an individual with a genetic mutation or developmental issues [3]. This is usually done with a biofluid sample such as serum or urine. The main applications of metabolic profiling are toxicity assessment of various environmental contaminants [4], biomarker and

drug discovery for human disease diagnosis and treatment [5, 6], nutrition [7], infections [8], and functional genomics [9].

1.4 Single nucleotide polymorphisms

The linking of a genomic variant with a phenotypic trait (functional genomics), often used in plants or bacteria, is similar to approaches involving single nucleotide polymorphisms (SNPs) in humans. Indeed, functional genomics is useful for determining the phenotype linked with a genetic mutation or variant, such as the impact of genetic heritability on cytokine production in the human immune response [10]. SNPs are mutations of one nucleotide in DNA strands in a minority of the human population compared to the majority nucleotide, but present in at least 1% of the population.

SNPs can also be used to predict the function of unknown genes by comparing the metabolic profile with that of a known genetic perturbation. [11]. Additionally, these nucleotide variants, for example a G (guanine) instead of an A (adenine), can be predictive and linked to certain diseases. An example of this is the E4 allele and its link to Alzheimer disease, affecting the apoϵ4 protein, which leads to a higher risk of early onset dementia [12].

Most SNPs occur in non-coding regions of the genome, but some do appear in coding regions and therefore may directly affect the gene sequence, expression, and/or gene product by producing a different amino acid. It is more probable for a given SNP to result in a deleterious or neutral effect than an increase in gene expression or enhanced enzyme activity: it is easier to break or do nothing than to improve. Due to the redundancy of the genetic code, some nucleotide mutations can have no effect on the resulting amino acid, often in the third position of the codon. Both GCA and GCG code for the alanine amino acid, meaning if the A mutates to a G or inversely, alanine will be added to the protein regardless. Other mutations will introduce an early stop codon, resulting in a truncated protein

which is either degraded or non-functional but still produced. SNPs occurring in non-coding regions can have effects on regulation by modifying transcription factor binding sites, chromatin folding, epigenetic modifications, enhancers, non translated RNA genes, and likely many more we have yet to discover. Indeed, non-coding regions of the genome are understudied due to the less direct link they have with the phenotype, while constituting 99% of the human genome [13], meaning many effects of SNPs are best studied by linking observable phenotypes such as metabolic profiles with the genotypic mutation.

A Genome-Wide Association Study (GWAS) is the observation of genomic variations with phenotypic traits across many individuals. It is based on the statistical testing of many variations to find the statistically significant traits associated with them [14]. For example, one of the first GWAS linked age-related macular degeneration, resulting in blurred vision, with two SNPs. GWAS can reveal links between a genetic variation and a disease without necessarily understanding why or how the two are linked, and can thus help in diagnosing or prescreening diseases without obvious symptoms. It can also orient research towards genetic targets that were previously not thought to be associated with the disease, and it can be used in epidemiology to discern group differences in response to diseases.

In 2008, GWAS was applied to metabolomics data, combining the genotypic variation associations with metabolic traits such as the concentration of a given metabolite in blood to form a metabolite genome-wide association study (mGWAS) [15]. This study tested the link between 363 serum metabolites and SNPs in 284 individuals, and demonstrated the concept of the ”genetically determined metabotype”. Many mGWAS cohort studies focus specifically on SNPs involved in or near enzyme-coding regions since those are the SNPs most likely to have a measurable effect on metabolism. Applications of mGWAS can be seen across a variety of fields, describing links between genetic factors and diseases, such as cancer [16] and kidney disease [17], or in plant physiology in

seed development for example [18]. mGWAS can thus be used with large-scale metabolite measurement approaches to produce entire profiles of metabolites across many individuals and detect significant shifts in concentrations associated with SNPs in those individuals.

1.5 Omics approaches for metabolic phenotyping

The "omics" sciences are branches of science focused on the comprehensive and broad study of the constituents within cells, tissues, or otherwise biological samples. The names of each branch end in the suffix "–omics" and each branch has its associated "-ome", such as genomics with the set of all genes being the genome. By describing and measuring the complete sets of molecules and biochemical processes that contribute to the survival and development of cells, tissues, and organisms, information can be gained, specific to an individual, condition or even timepoint, as well as entire populations. This of course generates large amounts of data relevant to each part of the cellular process. Each approach examines specific parts of biology, and the major approaches can be described as the following:

Genomics represents the sequencing of the genome of an organism and the study of its genetic or epigenetic sequence. It focuses on analysing all genes and their relationships to identify how they interact to affect the development of an organism as well as how it reacts to external changes. In human health, these molecular mechanisms are studied in relation to diseases and environmental factors, and more specifically in relation to this thesis, enzyme-coding genes are identified to link genetic expression mechanisms with translation of mRNA to enzymes. Depending on gene annotations available for a given organism, predictions of putative functions can be evaluated as regulatory, metabolic, or otherwise involved in cellular processes.

Transcriptomics is inherently closely linked to genomics, since the range of

possible transcripts is dependent on the genes present in the genome and their expression. While the genome of an individual is established, the transcriptome can differ depending on the tissue, conditions, and other factors. This is why transcriptomics, the study of mRNA produced by the transcription of genes, is a good indicator of gene expression and can unveil this first level of regulational difference between genome and transcriptome, but depends greatly on the location of the sample.

Proteomics draws the final link between genes and their products, as mRNA expression is not always correlated with enzyme translation and activity. By identifying and quantifying proteins directly, it indicates whether a gene was transcribed into mRNA and then translated into its final form. In the case of metabolic genes, often the products are enzymes or proteins directly involved in the transformation of metabolites.

Metabolomics involves the identification and sometimes quantification (often relative) of metabolites in a sample. Metabolites are small molecules used by enzymes in biochemical reactions. Metabolites found in an organism's metabolome (set of metabolites in a sample) can be either naturally produced by that organism (endogenous) or originate from external sources (exogenous) non-natural to the organism. By studying the link between metabolites and enzymes, hypotheses can be drawn on the reality of enzyme activities, as well as on entire metabolic processes and markers of specific conditions or changes. More specific branches of metabolomics include lipidomics, which is the study of lipids and their structure and function within the cell or organism. Lipidomics became a separate field of study due to the complexity and diversity of lipid structures and unique functions [19].

Finally, **fluxomics** consists of the measurement of the rate of intracellular fluxes of cellular metabolism. Fluxes are measured as the concentration of produced matter over time; in this case the matter consists of metabolites. It provides information on the dynamic biological processes that take place

within metabolism. Essentially, it combines the effect of all the previous omics approaches into one, in contrast to each approach only focusing on their own elements [20]. It is closely linked to a field of systems biology involving simulating these fluxes with the goal of explaining the complex biological systems behind metabolism.

There are of course various regulation points in between each omics approach, each of which adds a level of abstraction from the resulting phenotype. Indeed, a gene can exist without being expressed, and an enzyme can be produced without being active in reality. This can be due to regulatory mechanisms such as post-translational modifications, transcriptional regulation, epigenetics, and enzymatic inhibition, to cite a few.

1.5.1 Metabolomics

As mentioned briefly before, metabolomics is an approach aiming at the measurement and analysis of metabolites from biological samples. The major innovation of metabolomics is the global overview it provides of the metabolic processes present in a system, similar to older approaches like genomics, used to sequence entire genomes rather than focusing on a specific set of genes of interest. Metabolites and reactions take part in many different aspects of an organism's metabolism, which can be defined and split up into functional metabolic pathways such as glycolysis, and they can be more or less specific to an organ, tissue, cell type, or even cell compartment. A general metabolomics workflow is shown in Figure 1.

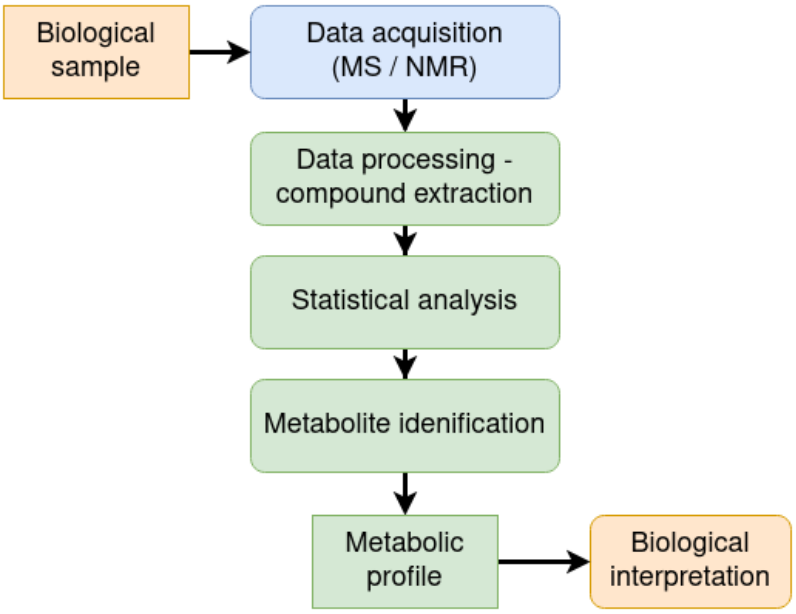


Figure 1: A metabolomics workflow, from the biological sample input to interpreting the data contextually.

The identification and quantification of metabolites has applications in various fields of research, such as studying the effect of the environment on organisms [21], biomarker discovery [22, 23], disease diagnostics [24], drug discovery [23, 25, 26], and food and nutrition research [27]. By acquiring and analysing snapshots of the metabolome at different time points, conditions, tissues, or individuals, biochemical effects and risks can be understood.

Metabolomics approaches are divided into two categories: targeted and untargeted. Targeted approaches focus on one metabolic pathway or class of compounds which defines the specific group(s) of metabolites to be analysed. This usually involves the addition of stable isotopic standards that are easily detectable when taking them into account as controls. These analyses often consist of tailored steps for certain chemical classes (e.g. amino acids, nucleotides…). Targeted methods are advantageous in their quantitative precision but lack breadth of coverage and detection capabilities, since they rely on prior knowledge and experimental design.

Untargeted approaches are oriented towards a global view of metabolic

fluctuations, usually in response to a given perturbation. These perturbations can be diseases, genetic, or environmental. Generally, untargeted analyses are used for hypothesis generation, which is then followed by targeted profiling on metabolites of interest. The metabolites of interest can then be quantified and analysed more confidently and thoroughly [28].

The two approaches can be combined by performing an untargeted acquisition of data followed by analysing data in a targeted way. It has the advantage of allowing data to be processed again, this time with a new perspective, without requiring the analysis of samples a second time.

The two main methods of measuring metabolites from a sample are mass spectrometry (MS) and nuclear magnetic resonance (NMR). Each analytical setup has the ability to detect a portion of the metabolome depending on the physico-chemical properties of molecules (e.g. polarity) [29].

1.5.2 Mass Spectrometry

MS can detect and identify metabolites with high sensitivity using their mass and charge, and can cover many different classes of metabolites due to the different methods available with different coverages. It consists of two technical steps: ionisation of metabolites followed by separation according to their mass-to-charge ratio. The signal produced by this method is then analysed to extract data from the resulting spectra.

Ionisation is the process by which a molecule (or single atom) gains or loses charges via the gain or loss of a proton or electron, often accompanied by other chemical changes. This results in a molecule, now known as an ion, with a negative or positive charge. In an MS procedure such as Electron ionisation for example, the sample can be bombarded with a beam of electrons which can ionise it or even fragment it into multiple smaller ion fragments. Different methods of ionisation are used depending on sample phase (liquid, solid, gas), sample size

or even sample salinity.

Among the various MS ionisation methods, atmospheric pressure chemical ionisation (APCI) is a gas-phase ionisation method used to detect medium and low polarity thermally-stable compounds (e.g. lipids), and is commonly used for trace analysis detection such as pesticides and drug metabolites. This is complementary to other methods like electrospray ionisation (ESI) which is better for high polarity metabolite detection and larger compound masses. Different methods cannot ionise certain molecules, such as volatile or thermosensitive compounds, meaning an absence of a metabolite in the resulting detection does not necessarily imply it was not present in the sample.

Compounds are then separated based on their mass-to-charge ratio by an analyser. Each molecule will have a different ion trajectory and speed which are generally characteristic of its behaviour, but some molecules are not able to be separated, resulting in identification ambiguity. They are then detected thanks to their charged nature, and the signal is converted to a mass spectrum.

Furthermore, MS techniques are often coupled with other methods to enhance coverage and precision. Many metabolites can correspond to the same mass or chemical formula, which then need to be separated in order to identify each metabolite correctly. For example, MS coupled with liquid chromatography (LC-MS/MS), combines both the physical separation provided by the LC, based on each metabolite's affinity for the phase(s) in the column, with the mass analysis from MS. MS is also commonly combined with a second MS to separate and detect fragments of the ions from the first MS. This approach, known as tandem mass spectrometry or MS/MS, makes it possible to overcome the issue of separating ions with very similar mass-to-charge ratios. MS/MS is also an essential tool for going further in metabolite identification due to the fact that the spectra can be compared with existing spectra in databases. These databases can be shared with the community [30], meaning that identifying metabolites using this method does not rely on internal identification databases.

1.5.3 Nuclear Magnetic Resonance

As opposed to MS, spectroscopy by nuclear magnetic resonance (NMR) often does not require chemical manipulation or destruction of a sample. NMR is a different metabolomics approach, and is based on detecting the shift in resonance frequencies of certain nuclei when exposed to an external magnetic field. This provides information on atoms as well as the neighbouring atoms and the bonds between them, which, combined together, provides enough information to more precisely identify metabolites than MS. It also is not limited by the physico-chemical properties of metabolites, meaning it can be used to measure a much wider range of molecules, and it is better for direct quantification, where MS is often best used for relative quantification. However, the major drawback of NMR is its lower sensitivity, meaning a minimum concentration is required in the sample for a compound to be detected, as well as the necessity of a larger base sample. For quantification, NMR does not require the use of standards which are often expensive to obtain, and is especially adapted for targeted quantification, as well as untargeted non-quantitative analysis.

1.5.4 Lack of metabolome coverage

Even by combining all possible analytical platforms, it is for now impossible to detect all metabolites with high confidence (see [31] for metabolic coverage assessment of MS data). This means that when preparing a metabolomics experiment, it is important to select the correct setup for the metabolites of interest beforehand. Due to the fundamental nature of metabolites as a molecular class, no one instrument can reliably measure every compound considered as a metabolite [32]. Metabolites are an incredibly broad range of molecules with different chemical properties and are present in varying concentrations in samples. This is in stark contrast with any gene or RNA related quantification: nucleotides are a very limited pool of molecules with similar

properties, meaning only one technique is required to even detect them. Multiple combinations of instruments are therefore often required to cover a broader variety of metabolites.

1.5.5 Downstream metabolomics analysis

Compound identification is generally preceded by statistical evaluation and analysis, often involving multivariate and univariate analyses, to categorise and predict sample properties, and identify major trends among the data. Multivariate analyses such as principal component analysis (PCA) and partial least squares discriminant analysis (PLS-DA) reduce the dimensionality of metabolomics data and can be used in classification, regression and prediction. These methods help select variables which could be of interest for identification.

Regardless of the metabolite identification method used, raw metabolomics data results in unidentified features in the form of spectrum peaks. Due to the nature of metabolomics experiments, once detected, measured features must be identified to ensure they are metabolites of interest. In order to be confident in an identification, the measurement must be compared and confirmed with the corresponding standard of that molecule. Often this reference standard is not available for new metabolites, meaning that new standards have to be bought in order to detect previously unknown metabolites. This leaves gaps in observations where an absence of measurement and identification cannot be equated to an absence of the metabolite in the sample.

Metabolite identification has been standardised in the community to improve communication and reuse of information. A system of four identification levels was proposed and is now widely used [33]. Since then, other initiatives have been developed [34], but the system proposed by the MSI (Metabolomics Standards Initiative) remains the most widespread. It is described as the following:

1. Identified compounds: the requirements for a level 1 identification involve

the use of a pure solution of the molecule (a standard) in identical experimental conditions as well as documentation of the spectral matching process used.

2. Putatively annotated compounds: a molecule is annotated as level 2 when the identification is based on spectral or chemical similarity with public spectral libraries.

3. Putatively characterised compound classes: a class is assigned to a molecule based on its phyisco-chemical properties or spectral similarity to compounds of that class.

4. Unknown compounds: a compound which can be differentiated and quantified but not identified or classified.

Prior knowledge of which molecules should be identified is essential when designing a metabolomics experiment due to the difficulty of obtaining a pure standard for many metabolites. Identifying novel metabolites remains difficult, especially when carrying out exploratory experiments, and reaching level 1 identification on all detected spectra is impossible. This results in a loss of information at multiple levels, and the resulting metabolic profile is not wholly representative of the sample.

1.6 Conclusion: a need to fill the gaps in metabolic profiling

Measuring metabolites has impacts in many fields and studies, such as human health, and one of its advantages, especially in the case of health-related studies, is that it is very close to the actual phenotype of the organism compared with transcriptomics. This results in cellular metabolic mechanisms that can be closely correlated with observable phenomena like diseases, and is why metabolic profiling can be very performant in diagnosing diseases. Metabolomics holds its appeal primarily because it captures the most dynamic

representation of phenotype and medical conditions. Metabolites are viewed as falling downstream of genetic, transcriptomic, proteomic and environmental variation [32].

The fundamental advantage of recent advances in metabolomics is the ability to measure large numbers of metabolites at once, going from metabolite-level analysis to metabolic profile-level analysis. This transition from single biomarker quantification to lists of hundreds of metabolites not only creates a need for downstream methods able to analyse these large lists of metabolites but also approaches involving prior analysis and selection of classes of metabolites to measure for future experiments. Knowing when and what to focus experiments on is not only essential for reducing costs and time spent, but also for more directed hypothesis generation as well as easier down-stream analysis.

In addition to this, the loss of metabolites along the way means that the resulting metabolic profile captures only a partial view of the metabolome in the sample, inherently misrepresenting the metabolic state. Being able to complete the experimental view of the metabolome with predicted metabolites of potential interest can add new information to the current analysis as well as improve future experimental design.

2 Metabolic modelling

The ultimate goal of metabolomics and biology in general is the full comprehension of every biological system in any given scenario. Ideally, this would serve to improve our understanding of the relationships between our bodies and everything they encounter, as well as why sometimes things go wrong. Of course, this is not (yet) possible, and creating models to represent and study reality is a step towards even better models and of course complete understanding. This ”classic” view of modelling helps understand and predict complex functionalities that are understudied.

In this thesis, the aim is to move past this use of modelling with the goal of predicting outputs that could be observable in a given condition. By developing the prediction of metabolic profiles using *in silico* methods, this can aid in the design of experimental studies and improve our knowledge of metabolism. As mentioned in the previous sections, designing experiments is difficult, time-consuming and expensive, and results may not always be positive, which leads to more experimental design to confirm and develop results. Being able to narrow down future metabolic targets is an essential part of optimising hypothesis testing. This can be achieved by modelling metabolism at the level of cells or tissues.

2.1 Bioinformatics, a necessary tool in the era of systems biology

Bioinformatics is the hybrid approach of developing and using computational techniques to support other areas of scientific research involving biology. By manipulating and interpreting biological data, biological problems can be solved and hypotheses can be generated. This data can be produced experimentally, come from public cohort data, or other online databases. Bioinformatics has been used since the 1950s, becoming essential when comparing sequences of genes manually became impractical. As the production of biological data becomes easier, cheaper, more widespread and on a larger scale, techniques must be developed to analyse it in order to keep up.

Beyond analysing and comparing biological data, entire new fields of bioinformatics have flourished, giving rise to predictive tools in an area of study known as systems biology. By combining pieces together, as opposed to taking pieces apart and looking at them individually, systems biology gives a view of the larger picture and this enables the modelling of complex biological systems at various levels (organism, tissue, cell…). The hypothesis behind systems biology

is that this combination of interactions between individual systems possesses additional properties as a whole (holistic), resulting in emergent properties. One of these complex biological systems is metabolism, which can be especially well simulated by a model due to its interconnected nature.

2.2 Modelling principles and goals

Modelling entire complex systems is based on the simplification of what is known to a certain degree where the model remains a valid approximation. This is done by removing parameters deemed with minimal impact on the results, or those difficult to measure experimentally and therefore complicated to validate biologically. This approximation remains valid as long as there is enough information left in the model to result in realistic predictions. Generally, a model is defined for a given objective, meaning a model is developed and used to answer a predefined question, or to generate questions and hypotheses. In this sense, it is a specialised model which should only be used for the corresponding topic, and should be re-evaluated if used for other predictive purposes. The goal is to obtain a scientifically accurate prediction which can then be reflected back onto reality, but no one model describes an absolute truth.

Scientific modelling encompasses many different types of models, such as conceptual models, graphical models, mathematical models (using mathematical concepts: statistics, game theory...), and computational models (more algorithm and simulation focused). Computational models are often used to simulate a complex system using a mechanistic approach. They aim to replace manual and "intuitive" approaches with a more parameter-centric computer-based experimentation.

Understanding genes and gene expression is essential to the enrichment of the community's knowledge of metabolism, and more specifically the biological processes involved in getting from a gene to an enzyme and finally to an

active biochemical reaction. Ultimately, it is by studying metabolism and metabolic-related mechanisms that we can understand the final effects of gene regulation on metabolism, due to the many intermediate steps in between. Biologists can work backwards from metabolite measurements to understand how reactions are linked together, which is of course a necessary preliminary step for fundamental biology and developing any sort of model representing reality. Now that these models exist, they can be improved upon with additional knowledge gained from experiments, and, more relevant to this thesis, can be used to generate this new knowledge and create a feedback loop of model improvement.

We cannot predict the effect of a blocked enzyme on the entirety of metabolism by looking at how reactions are linked together. Just because an enzyme is repressed does not mean that the substrate metabolite will accumulate in the cell, as other compensatory mechanisms may exist and only activate in the event of an abundance of that metabolite. Being able to simulate metabolite fluctuations is essential to understanding the propagation of disruptions across an organism's metabolism.

2.3 Metabolic modelling methods

There are a multitude of networks developed and used for biological applications, such as protein-protein interaction networks, cellular signalling networks, between-species interaction (trophic) networks, and many more. When applied to metabolism, networks consist of metabolites, usually unique single entities, linked by reactions. The simplest form of representing a reaction is by combining an enzyme with its substrates and products as well as stoichiometry information. The next level is the combination of multiple reactions and metabolites contributing to a cellular function, known as a pathway. Pathways can be defined differently depending on the method of

segmentation, and can overlap, sharing metabolites, or not overlap. As a whole, these various levels of an organism's metabolism are available in databases, containing all known information on each entity and their relationships. This information can then be harnessed by using either graphs, for a more static and descriptive approach, or constraint-based modelling, for more predictive and quantitative methods.

When modelling metabolism, the model can be structured in different ways depending on the biological question and technical solutions available. Since a model by definition is a reduction of information, choosing which information to keep and which to remove is essential when designing or selecting the method of modelling. It is also dependent on the data available for a given process: if a parameter is impossible or difficult to measure, then the model can be based on a hypothesis to either remove the parameter or estimate it theoretically.

2.3.1 Databases, gathering knowledge on metabolism

Metabolic databases include information on each metabolite, reaction, enzyme and gene known for a given organism, as well as how they are linked together. Kyoto Encyclopedia of Genes and Genomes (KEGG) [35, 36, 37], is a Japanese collection of databases containing many aspects of biology including, but not limited to, genomes, metabolites, biological pathways, and diseases. It is a comprehensive representation of the interactions between molecules, reactions and genes in the form of pathways, and chooses to represent pathways as visually simple manually laid out maps. MetaCyc [38, 39, 40] is a metabolic pathways- and enzymes-centric database, and is viewed as an extensive online encyclopaedia of metabolism. Many pathways and enzymes are backed by mini reviews and other literature references, and it also includes data on metabolites. Reactome [41] is another online database of biological pathways, with several dedicated organism-specific databases. It focuses on the visual

aspects of representing biological pathways as well as sharing the data in a computationally accessible format. Rhea [42] is a manually curated database, specialising in biochemical reactions.

In the databases containing pathways, each pathway is defined and named differently and can vary in size compared with the equivalent in other databases [43, 44]. Some pathways exist in certain databases and are merged or non-existent in others. This leads to issues mapping between databases, and requires a choice of database when using pathway data to enrich metabolomics results.

Most databases use identifiers for reactions, genes and metabolites. Gene identifiers are generally well-standardised across multiple databases, with common identifiers including Ensembl [45], NCBI Entrez [46], and even the basic gene symbols are well recognised. Enzymes have a universal enzyme class, also known as the Enzyme Commission (EC) number [47], which classifies enzymes according to the reaction(s) they catalyse, meaning multiple enzymes can have the same EC number if they carry out the same metabolic function. By contrast, UniProt [48] identifiers are uniquely assigned to enzymes, which contains all annotated information as well as the link to genes that code for the enzyme. Finally, metabolites have different identifiers such as Chemical Entities of Biological Interest (ChEBI) [49], PubChem [50], and Human Metabolome DataBase (HMDB) [51], which serve to assign a unique identifier to each metabolite to help find, archive, and consolidate diverse metabolite names. A different type of identifier is based on the chemical structure of metabolites, such as SMILES [52] and InChi [53]. These identifiers also have the goal of being unique: some rare cases can occur where an identifier corresponds to multiple metabolites but these tend to be fixed quickly. They encode chemical properties (molecular structure, stereochemistry...) into the identifiers themselves, meaning that information can be extracted and used without querying external databases. Despite these identifier development efforts and pushes towards a more consistent approach to metabolite nomenclature [54], no real homogeneous

naming convention has been chosen as the consensus yet. For metabolites there is a distinct lack of combined effort for harmonisation of identifiers, which leads to issues such as misidentification of metabolites, difficulty in mapping from names to ChEBI or PubChem, and the necessity of manual curation to check identifiers. Common names, which can differ from a missing hyphen to a completely different chemical description for the same molecule, need to be associated with at least one standard identifier upon identification, but this is not always the case when producing metabolomics data. For example, 20alpha-Hydroxyprogesterone is also called 20a-Dihydroprogesterone. When inputting 20alpha-Hydroxyprogesterone into identifier mapping services such as MetaboAnalyst [55], 20a-Dihydroprogesterone is returned, but if the hyphen is removed in the input, many different metabolites are proposed as equal matches. Finally, this molecule has many different synonyms listed on HMDB such as 20alpha-Progerol and 4-Pregnen-3-one-20alpha-ol which both refer to the same compound.

Some of the previously mentioned databases also provide their information through visual networks of connected nodes and edges. KEGG for example displays pathway-based networks with metabolites as nodes and reactions linking them as edges. Such networks contain the complete known set of metabolic and otherwise metabolically linked chemical or physical processes. Importantly, these are not solely lists of reactions but include how every object is linked to one another, which is the defining feature of a network. In principle, this represents discrete or binary data, meaning that either two entities are linked or not.

Other large metabolic databases include BiGG [56], MetaNetX [57], BioModels [58], and Metabolic Atlas [59, 60, 61], all four of which are intrinsically linked to systems biology and storing metabolic information in a suitable format. The main format for storing a metabolic network is Systems Biology Markup Language (SBML), an XML-based data format for storing and describing models in biology

[62, 63, 64]. Models in the SBML format include identifiers for various databases depending on the level of curation and annotation, which is important when mapping to or from biological data.

BiGG is an open-source knowledgebase of more than 70 genome-scale metabolic network reconstructions with standardised BiGG IDs. This allows the user to quickly obtain the ID from one model's metabolites, genes or reactions in another model, even across species. MetaNetX is a similar repository for many metabolic networks with their own "MNXref" IDs for metabolites, genes and reactions, as well as external database links when available. BioModels is an open repository for mathematical models curated and hosted by the EBI (European Bioinformatics Institute). Metabolic Atlas is specifically built to browse and view a set of genome-scale metabolic models (GEMs) made with the goal of being the community's consensus models for 7 species (human, yeast, fruitfly, mouse, rat, worm, and zebrafish). Since these models can be large, it can be useful to browse the model's entities online to understand each entity's properties, such as the genes responsible for a given reaction.

All of this data contained within databases is of use when constructing models that represent various aspects of biology, including metabolism. By translating the interconnected knowledge into a functional or structural network or model, new information can be extracted from what was already known, using the links annotated on each entity, as well as the entity's own data. The following sections describe these metabolic networks that are constructed from knowledge available in databases and in the literature.

2.3.2 Reconstructing metabolism to create genome-scale metabolic networks

Genome-scale metabolic networks (GSMNs) aim to encompass all known metabolic genes, reactions and metabolites as well as the interactions between

them for a given organism [65, 66, 67]. They are created "bottom-up" from genomic and literature data, in contrast to "top-down" approaches, which use large quantities of data to infer interactions and generally do not form functional models for mathematical modelling [67]. The reconstruction process starts with the genome sequence assembly for the chosen organism:

1. **Draft reconstruction**: from genome annotations, identify metabolic reactions and combine into a draft reconstruction with potential metabolic reactions.

2. **Refinement of reconstruction**: various checks and addition of information on substrates, cofactors, formula, stoichiometry, compartments, transport reactions...

In these networks, metabolites and reactions are represented as a stoichiometric matrix, containing information on the proportions of each metabolite involved in each reaction. For example, the combustion of methane (CH_4 with O_2) produces CO_2 and H_2O, but one unit of CH_4 contains 4 hydrogens and therefore produces 2 units of H_2O while consuming $2O_2$. The 2 units of O_2 and H_2O required to make the reaction balanced are the stoichiometric coefficients of the compounds for that reaction, as well as 1 for CH_4 and CO_2.

Enzymes are annotated by matching metabolic genes to their corresponding enzymes both structurally, using the gene sequence, and functionally, using databases and other information like EC numbers.

In GSMNs, genes are linked to reactions through their Gene Protein Reaction (GPR) rules (see Figure 2). In the case of an "AND" GPR, all of genes in the reaction's GPR must be active for it to be active, and in the case of an "OR" GPR, at least one gene in a reaction's GPR must be active for it to be active. This concept represents some aspects of gene regulation, requiring some genes to be active either because they code directly for the gene or are known to regulate the enzyme's expression directly. The "AND" GPR can model protein complexes

and how some enzymes require multiple proteins to join together to function. The "OR" GPR can model the gene duplication that can occur throughout evolutionary processes, resulting in multiple enzymes with the same function.

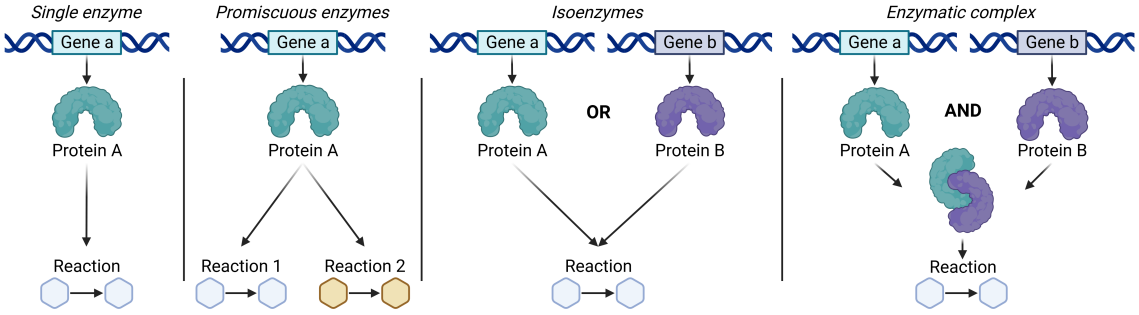


Figure 2: Gene-Protein-Reaction relationships. In metabolic models, isoenzymes correspond to the "OR" GPR, where at least one of the genes is required to be active, and enzymatic complexes correspond to the "AND" GPR, where multiple genes are required to be active for a reaction to take place. Figure based on the GPR figure in [68].

Homeostasis, or the internal balance of conditions, is maintained and regulated by three components: a receptor, a control centre and an effector. The conditions can range from body temperature to pH levels, concentrations of sodium to blood sugar levels, and must be kept within predefined ranges even when external changes affect the organism. Homeostasis is incredibly important for the function of many metabolic processes due to their specific environmental requirements. A major example of this is osmosis, the spontaneous movement of molecules through a membrane, which can only occur if the molecule is in a lower concentration in the destination compartment. There are transport mechanisms in place which transfer molecules in the opposite direction to maintain ideal concentrations on either side of the membrane. In GSMNs, metabolites can be separated into compartments such as cytosol, mitochondria, endoplasmic reticulum, and peroxisome etc., resulting in duplicated version of metabolites for each compartment. Metabolites are transported between each compartment via dedicated reactions converting a metabolite in one compartment to the same metabolite in the destination compartment. This transfer via transport reactions can occur between any two compartments in

contact with each other. A specific type of transport reaction is the exchange reaction, which transports metabolites out of the cell, creating an ”exchange” with the ”outside” of the cell. The ”outside” of the cell can model the cellular medium or a biofluid such as blood vessels connecting and supplying cells with nutrients as well as providing an output for waste and exports.

GSMNs are used in many applications such as maximising the production of certain metabolites (metabolic engineering) [69], studying the effect of drugs on pathogens [70], discovering enzyme functions [71], comparing the metabolism of different cells, tissues or organisms and predicting interactions between them [72], and studying disease states in relation to metabolism [73, 74].

2.3.3 History of metabolic network reconstruction

The first genome-scale metabolic networks were developed for bacteria due to the smaller genomes and easier to access information in 1995 for *Haemophilus influenzae* [75] followed by *Escherichia coli* [76]. Multicellular organisms are more difficult to reconstruct due to the size of the genome, less knowledge and the various cellular compartments. The first multicellular reconstruction was for *C. elegans* in 1998 [77]. Since then, many other organism models have been reconstructed, including updates to older versions.

After the publication of the complete human genome in 2004 [78], human metabolic network reconstructions started to be developed. Curated collections of biochemical reactions in human cells were built in 2004 (HumanCyc) [79] and 2005 (Reactome) [80, 41] but the first metabolic models of human metabolism were published in 2007 as Recon1 [81] and EHMN (Edinburgh human metabolic network) [82] with more unique metabolites and reactions that HumanCyc. Recon2 was the next step in human metabolic models [83], combining everything from EHMN, Recon1, and HepatoNet1 [84], as well as extra modules, and was then refined in 2015 with improved GPRs and other updates to form

Recon2.2 in 2016 [85]. In parallel, HMR [86] was built combining Recon1, EHMN, HepatoNet1, Reactome, HumanCyc, KEGG, and the Human Metabolic Atlas [87], and then further updated as HMR2 [88]. In 2017, another human metabolic model was developed called iHsa [89] as an extension of HMR2, combined with the rat metabolic model iRno. Recon3D [90] was then developed using Recon2 and HMR2 as well as other reaction sets. Finally, the most recent human metabolic network is known as Human1 [60] and is provided via github, meaning it can be regularly updated by the community with proper versioning and without having to release a new version of the model separately. For instance, I identified some GPR-related errors in the model, and the corrections were rapidly integrated into Human1 v1.15, after being reviewed and validated by other contributors.

2.3.4 Metabolic graphs to discover network structure and relationships between elements

A different approach to structuring information is using graphs and graph theory. Graph theory revolves around the relations between elements: how they are connected and how they are not connected. It has its origins in mathematics, more specifically in recreational maths problems, and is now used widely in many different fields of application, such as chemistry, computer science, maps, sociology, and even linguistics.

Graphs are useful for exploring structured knowledge thanks to the edges not only linking entities (nodes) but also describing the links with meaningful connections. This allows for automated association generation between many types of data, such as metabolic signatures and biomedical concepts [91], as well as hierarchical structures such as ontologies. Hierarchical ontologies are structured tree (more precisely directed acyclic graphs) representations of knowledge using controlled vocabulary as labels to describe objects. The most

well-known biological ontology is GO (Gene Ontology) [92, 93], which classifies genes into a tree structure with three main branches (Molecular Function, Celullar Component, Biological Process). An ontology based on small molecules is the ChEBI Ontology [49], which organises molecules such as metabolites based on their Molecular Structure and Role. The ChEBI ontology uses different relationships between nodes, with generic links like ”is a” and ”has part”, and more molecule-specific links such as ”is enantiomer of” etc.

Graph theory is an expansive field which uses various approaches to explore and utilise graphs. Algorithms have been created to calculate various properties of either the graphs themselves or the entities within them. A useful example is calculating the shortest path between two entities (nodes) in a graph. The shortest path is the path between two entities such that the sum of the weights of the edges along the path is minimised (when no weight or equal weight is defined, the shortest path is a path with the minimum number of edges). Many algorithms have been developed to solve this problem such as Dijkstra's algorithm, Bellman-Ford algorithm, and Floyd-Warshall algorithm. The shortest path length (or total weight), once calculated, can be considered as a distance between two entities. The graph's diameter can be determined by calculating all of the shortest paths between each pair of nodes and taking the length of the longest path.

In addition to the shortest path, many other metrics and algorithms exist to exploit graphs. Centrality is a measure of the importance of a node in the graph, which can be defined in different ways depending on the goal. It can be based on the number of edges connecting each node, or based on the distance to all other nodes. Assortativity describes the tendency for nodes to connect to other nodes with similar properties. The PageRank algorithm, well-known for its use by Google to rank search results, is another measure of importance based on the capacity of a node to be a key connecting point in the network.

Graphs are also extremely useful in visualising information as the simplicity

of the edge and node combination means that many different styles can be applied, and large scales can be reached. Many different libraries and tools exist for visualising a graph stored in a file, such as the igraph libraries (R, Python...), Cytoscape for general graph visualisation with special applications in bioinformatics [94], and MetExplore [95, 96] for specialised visualisation of metabolic networks as graphs.

Metabolic networks can be used to create graphs in multiple ways [97]. A reaction graph contains the nodes as reactions, with the edges representing a shared metabolite between reactions (the product of one is the substrate of the other) (left of Figure 3). A compound graph is the compound-centric counterpart: two nodes are connected by an edge if one is the substrate of a reaction and the other the product of the same reaction (middle of Figure 3). Bipartite graphs contain both types of nodes but the edges can only link different types: a reaction node can only be linked to metabolite nodes and vice versa (right of Figure 3). Choosing how to model metabolism using graphs depends on the biological question and available data (on metabolites or reactions), and can greatly change the interpretation and conclusions drawn from the graph.

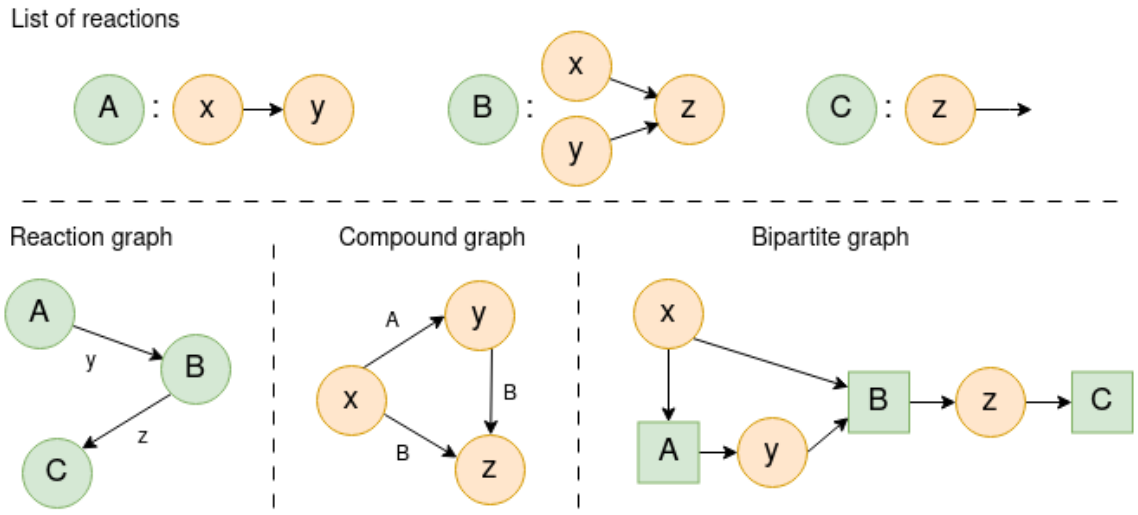


Figure 3: A metabolic network shown using different graph models.

Finally, graphs can also be directed or undirected (and even mixed), meaning

that the edges in the graph contain information of their directionality or not. Technically, all reactions are reversible, but an enzyme is required for most reactions to take place in biological conditions, leading to certain reactions only able to occur in the direction catalysed by the enzyme. For instance, a biochemical reaction can be irreversible, carrying out the metabolic transformation only in one direction, or reversible, able to go back and forth in the transformation. An irreversible reaction has a defined substrate(s) and product(s), and forms a point of no return for the path of metabolite transformation. This affects the calculation of shortest paths for example as a path may be possible from a source node to destination node in an undirected version of the graph, but in the directed equivalent the destination node could be unreachable, resulting in an "infinite" distance between the two.

Choices must therefore be made when defining a graph type and its properties from a metabolic network. For instance, reaction graphs are less useful for metabolomics data and simulating metabolism due to the implicit nature of how metabolites are described. Bipartite graphs are good for explicitly modelling both metabolites and reactions, while compound graphs provide the transformation information between two metabolites within the edges themselves. In general, metabolic graphs are directed as the irreversibility of reactions is important when considering modelling and biological coherence. Undirected metabolic graphs can still be useful for calculating paths between entities for example, which could reveal metabolites that are close when choosing to disregard the directions of reactions.

2.3.5 Quantitative and dynamic metabolic modelling

Overall, reactions can be viewed as chained together by the metabolites involved as substrates or products. One reaction's product is another reaction's substrate, and this links the two reactions, rendering them and their activities

dependent on one another. Most reactions are catalysed by an enzyme, a protein coded for by a given gene. Without these enzymes, the corresponding biochemical reactions would not be able to occur at rates fast enough to sustain life. This means that if a deleterious mutation occurs in an enzyme-coding gene, that enzyme will not function correctly and the reaction will most likely not take place spontaneously. In addition to metabolites as the main substrate or product of an enzyme, some enzymes require cofactors which can be inorganic (such as zinc) or organic, such as the coenzymes ATP, NADH, and NADPH which transfer chemical groups from one enzyme to another. These coenzymes are incredibly common to many different enzymes and are continuously regenerated to maintain a stable level inside the cell.

Enzyme activities are confined by the kinetics rules based on the chemical laws of molecular binding, transformation, and cleaving. Kinetics, or the study of the rate of enzymatic catalysis, can help understand the biochemical mechanisms involved in the reactions and their role in metabolism, by predicting the evolution of metabolite concentrations over time. It also includes the study of the control and regulation of their activities, based on regulatory mechanisms via inhibitors or activators and how they can affect the reaction rate. Once enzymatic kinetics parameters are known, they can be used to model metabolism using ordinary differential equations (ODE). This appears to be the most evident way of modelling metabolism: by using knowledge of how fast each reaction takes place combined with initial concentrations of metabolites, a kinetic model should be able to predict metabolite concentrations over time. The major issue with kinetic models is the need for kinetic enzyme parameters which are difficult to measure and not available for every enzyme in every condition, but they can be estimated using various parameter estimation methods [98, 99, 100, 101]. *In vitro* measurements can be used for single enzyme activities but they can cause unrealistic model behaviour when put in relation to one another and with regulatory mechanisms [102]. In general, kinetics data remains sparse and often

information is only available for a small number of well-known reactions and enzymes. These data limitations can result in uncertainties in predictions which render the simulations computationally unfeasible or expensive [103], which leads to these analyses being restricted to small networks and pathways.

The shift away from considering the temporal dimension of kinetics-based modelling was driven by the simplification of ODEs into linear equations, by imposing the steady-state hypothesis (explained below). A well-suited class of methods for this global metabolic modelling is constraint-based modelling (CBM). By using genome-scale metabolic networks, CBM can compute steady-state metabolic fluxes (the flow of metabolites) through biochemical reactions [104]. In contrast to methods requiring information on the kinetics of enzymes, CBM needs very little data of enzymatic parameters and metabolic concentrations to function. Networks for CBM play a central role in our understanding of the relation between genotype and phenotype due to its intermediary position. They are used not only to explain metabolic mechanisms but also predict various areas of metabolic processes.

The major requirement for creating a functional model from GSMNs is physiological data to compare the model predictions to, since evaluation and validation of the model relies on comparing predictions to reality. The two main steps are:

1. **Conversion of reconstruction into computable format**: convert network into a mathematical format and set objective function and constraints.

2. **Network evaluation**: tests based on mass and charge balance, gaps, dead-ends, blocked reaction and comparison with known properties.

These steps form a feedback loop of model improvement by using results from the evaluation step to change entities and parameters in the refinement of the model. CBM methods are based on mass-balance across the entire metabolic network, using both known reaction stoichiometry as well as the

main hypothesis of CBM: all internal metabolites are at a "stable" concentration and are therefore consumed as soon as they are produced by any reaction (in other words, their concentrations stay unchanged). This has the advantage of being able to solve problems with thousands of reactions, but its drawback is that it prevents the prediction of the internal concentrations of metabolites. The metabolic models used in CBM integrate not only reaction and metabolite data but also metabolic gene information and how the genes are related to reactions. The model is therefore a global representation of a given organism's metabolic knowledge, from genes to metabolites.

The fact that GSMNs contain information on both genes and proteins as well as metabolites provides a single platform to which transcriptomics, proteomics, and metabolomics data can be integrated, for a multi-omics approach to studying metabolism. This data and network, combined together, provide a specific simulation of metabolism in a given tissue [105]. These models, created as an "instance" of a network, can instead be tailored to a specific condition, such as different cellular environments, growth conditions or diseases such as cancer [106, 107, 108]. By modifying different parameters in the network such as exchange reactions and redefining the objective function (seen as the biological "goal" of the cell, like biomass production), a different model can be generated from this static network. This creates the key link between phenotype and gene variations, which is mechanistically more informative compared to standard GWAS data for example, which only provides significantly associated phenotypic traits with gene variations without an explanation of why these links exist [109]. Beyond modelling a single organism on a genome scale, networks can also be built to model a specific tissue or cell type [110] or even multiple tissues linked together [111]. Often, tissue-specific models are derived from generic models by removing genes that are not expressed in that tissue using transcriptomics data.

Fluxes exist under defined flux constraints, known as upper and lower

bounds, attributed to each reaction in the network. These constraints have default boundary values assigned by the network's creators which define both the range of possible fluxes and consequently reaction directionality (reversible if both negative and positive values are allowed, or irreversible if only positive or only negative values are possible). By changing these bounds, different metabolic states can be modelled, for instance the simulation of a complete knock-out (KO) of a reaction by setting both of its bounds to zero. By knocking out a gene and therefore its associated reactions in the model, an IEM can be simulated. The reduction of flux through a reaction or multiple reactions (knock-down) can also be simulated in the model, representing reduced enzyme activity due to some effect of treatment, regulation, or exposure to xenobiotics.

2.4 Limits of metabolic networks in the integration of regulatory mechanisms

Biologically, enzyme production and activity can be regulated in different ways. The first and more extreme method is through alternative splicing, which, through the different combinations of intron and exons from one gene, produces variations of the same enzyme with distinct functions and structures. This can be predicted using splicing prediction tools from the genetic sequence [112, 113], and can be measured using transcriptomics since splicing directly affects mRNA transcripts. Some splicing mechanisms can result in smaller proteins due to alternative splice sites [114]. Other genetic-related regulation mechanisms include gene expression modulation via transcription factors or chromatin remodelling, and mRNA regulatory processes, where different steps of the translation process can be regulated by initiation, elongation, and release factors. All of these genetic-level regulations can be affected by genetic mutations such as SNPs, for example in intronic or exonic regions, or disrupting transcription factor binding sites.

Cellular compartmentalisation plays a key role in how an enzyme functions. Enzymes are spatially separated by the cellular membranes of each cellular component, restricting the reach of enzymes to the substrates currently in their compartments. Certain enzymes sometimes require a specific pH and temperature to function, which can be found in the mitochondria for example, and the directionality of reactions can also change depending on pH.

Enzyme activity can also be directly influenced by their own substrates and products in positive or negative feedback loops. The concentrations of substrates are highly controlled and some enzymes are only activated if there is a minimum level of its substrate nearby, even if the enzymes are otherwise available. Of course, no enzyme can function without its substrates, which makes it intrinsically linked to the flow of metabolites through other reactions located up and downstream. More specific enzyme regulation mechanisms include allosteric regulation, meaning the enzyme has a second binding site to which a ligand can bind and modulate its activity. These ligands are molecules which can either affect the enzyme positively (activation) or inhibit its activity. For example, a well-known allosteric enzyme is phosphofructokinase-1 (PFK-1) [115], as it is an important enzyme of glycolysis. It is regulated by many inhibitors, such as a high ratio of ATP to ADP inhibiting PFK-1 and therefore glycolysis, and activators, such as fructose 2,6-biphosphate in eukaryotes, a feedforward stimulation resulting in an acceleration of glycolysis when glucose is abundant.

Enzymes can also be regulated by controlling their degradation rate, directly affecting the number of enzymes available to catalyse reactions. Proteasomes, which are protein complexes located in the nucleus and cytoplasm, degrade unneeded or damaged proteins and enzymes, after being tagged with a small protein called ubiquitin, by breaking their peptide bonds.

In GSMNs, some enzymatic regulatory mechanisms can be modelled while others are not included in the network. Any direct and constant effect on a gene or enzyme can be modelled by affecting the parameters of the corresponding

reaction, such as rendering it completely non-functional. Regulation via substrate availability is baked into how CBM models simulate fluxes: if the previous reaction is not producing a metabolite in a given state, the following reaction using that metabolite as a substrate will not be able to function. Compartmentalisation is also included in GSMNs as metabolites have different versions for each compartment, leading to compartment-specific reactions.

However, detailed regulation like allosteric regulation is not modelled by GSMNs due to its relation to time-dependency: this sort of regulation fluctuates over time which is difficult to model on a genome scale because of the quantity of information that would be required. Additionally, enzyme degradation is not modelled in GSMNs as enzymes are not entities that can be created and degraded over time.

2.5 Flux simulation using CBM

CBM can be used to predict fluxes, or the flow of metabolites through a reaction, at steady state under various conditions. This is achieved by defining metabolism as a system of linear equations, which represent the mass balance equations of metabolites, composed of the reaction flux vectors involving each metabolite. Metabolites are the links between reactions, meaning that when a metabolite is "produced" by a reaction, it is immediately "consumed" by the next reaction.

First, reactions and metabolites are modelled using a stoichiometric matrix (S) with metabolites as rows and reactions as columns, and the values as the coefficients for each metabolite's involvement in each reaction. This results in a generally sparse matrix since most reactions only involve a few metabolites. Figure 4 shows how each constraint is added to the model and the resulting equations detailed in this section.

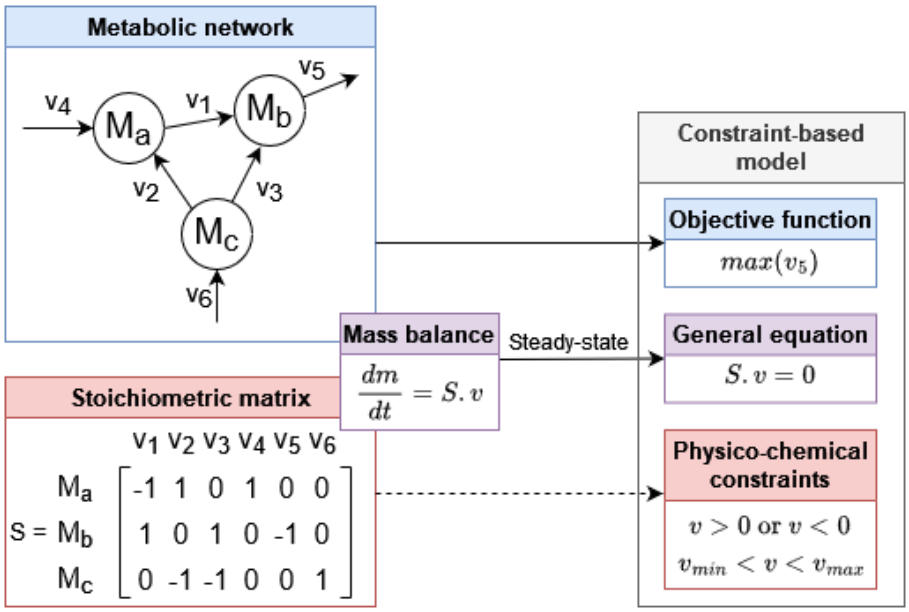


Figure 4: Constraint-based modelling constraints are defined using model assumptions. The objective function constraint is used in FBA for example.

Mass balance equations can then be mathematically defined for each intracellular metabolite, describing the evolution over time of each metabolite m_i:

$$\frac{dm}{dt} = S.v \tag{1}$$

where $m = (m_1, m_2, ..., m_i)$ is the vector of concentrations of intracellular metabolites, and $v = (v_1, v_2, ..., v_n)$ is the flux vector.

CBM relies on one major assumption, the first of which is the general pseudo steady-state of the model's internal metabolites. This assumption is based on the fact that internal dynamics happen quasi-instantaneously, meaning that the model reaches the steady state instantly and that its transient behaviour should not be taken into account. Consequently, at steady-state there is no accumulation or depletion of internal metabolites, which means that the rate of production is equal to the rate of consumption for each metabolite [116]. This can then be

expressed mathematically as:

$$S.v = 0 \tag{2}$$

Even with this steady-state assumption, the possible space of all solutions is undetermined because the number of unknowns (i.e., reactions (n)) is usually greater than the number of equations (i.e., metabolites (i)). Additional constraints are added using the thermodynamic and capacity properties of reactions when this information is available: irreversibility and reversibility, and maximum uptake or measured flux values. This is done by adding minimum and maximum bounds for the fluxes of reactions: $v_{min} < v < v_{max}$, as well as setting both bounds to either only positive or only negative values for irreversible reactions: $v > 0$ or $v < 0$. These constraints define a range of feasible values in the solution space but no unique solution [117].

Flux Balance Analysis (FBA) is a CBM approach which relies on one additional assumption in order to restrict the solution space further: the optimisation of the cell through a biological goal, or objective function (Bf). This means that when modelling using this method, the solution obtained is optimised for a certain metabolic objective [117] like biomass production. The addition of this optimisation problem means that the system becomes solvable using linear programming methodologies, but only provides one solution among many possible solutions. The final maximisation problem can be stated as:

$max(v_{Bf})$ where

$$\begin{cases} S.v = 0 \\ v > 0 \text{ or } v < 0 \text{ for irreversible reactions} \\ v_{min} < v < v_{max} \text{ for enzyme capacities and known flux limits} \end{cases} \tag{3}$$

Defining the objective function is essential to the correct simulation of

the metabolic state of the cell. The objective function should closely match the cellular metabolism's "goal" for a given state. The typical assumption is that cells have evolved to optimise growth, or the creation of biomass. Other objective functions include optimising ATP production, for the most energy efficient metabolism, the production of a specific metabolite, or a maintenance function, simulating cellular survival without growth.

A consequence of the steady-state hypothesis is that internal reaction fluxes cannot be used to simulate the quantity of any single metabolite. Conversely, reactions which import and export metabolites to and from the model, called exchange reactions, are able to be used to somewhat "quantify" the amount of that metabolite being exported to the extracellular space. This "quantification" is limited by the initial flux bounds set on not only the exchange reaction but also the other reactions in the network, and does not represent a real amount of that metabolite in the extracellular space. Despite this, this exchange reaction-based flux simulation is how we are able to draw a parallel between the model's simulations and experimental metabolic profiles (see Figure 6). By simulating fluxes representing metabolite export and import to the cell, metabolic profiles can be predicted representing how much a metabolite's export rate has changed between two conditions, for example a healthy condition and a disease condition (see Figure 7 for more details).

Thanks to the steady-state assumption in combination with other assumptions (such as the objective function in FBA) to reduce the solution space, CBM methods can simulate fluxes using no kinetic parameters or metabolite concentrations. Various methods exist to explore this constrained solution space, with FBA using stoichiometric and initial model constraints in combination with the optimisation of the objective function to provide an (one) optimal solution.

As mentioned previously, there are not enough constraints to define one single optimal solution, meaning that there are many different combinations of flux values for each reaction which can satisfy the constraints. Different methods exist

to explore the solution space and its possible combinations, two of which are Flux Variability Analysis (FVA) and random sampling (detailed in the following section). FVA can describe the extent of the solution space by providing the minimum and maximum possible values each reaction's fluxes can take in any of the viable solutions. This represents the extreme possibilities of a reaction, i.e. the most extreme cases that solve the flux equations. Therefore, FVA reports two possible solutions (the minimum and maximum values), leaving the rest of the flux solutions to fall in the interval between the extremes. The behaviour of fluxes within these boundaries remains undefined by FVA.

2.6 Random sampling in metabolic networks

Random sampling is a statistical method where each sample has an equal probability of being drawn. It ensures the chosen samples are an unbiased representation of the source population. It is often used when generalising predictions and inferences about a particular population, removing the need to collect data from every individual in that population.

Random sampling the fluxes of each GSMN reaction within the solution space brings us closer to the real distribution of fluxes which satisfy the model constraints. Any one possible solution is a specific combination of flux values for each reaction, and some flux values are more frequent than others, meaning that they appear more often in different valid solutions. These many solutions can therefore outline each reaction's flux distribution and thus reveal the most frequent fluxes along with the variability of the values, including the extreme flux values like with FVA. Some of the first applications of random sampling in CBM were to determine the effects of enzymopathies on red blood cells [118] and to study the impact of diabetes, ischemia and diets on human mitochondria [119].

Figure 5 highlights both FVA and random sampling, two CBM methods

for assessing the variability of flux values, in two different conditions (WT and MUT). FVA predicts two extreme flux solutions (minimum and maximum flux bounds) for each reaction in the network (Figure 5B) under the predefined constraints, while sampling calculates many different solutions spanning the solution space (Figure 5C). The FVA interval represents the range of possible values that the fluxes of a given reaction can take. In an extreme case, like a complete KO or a single flux value, both the upper and lower bounds of the involved reaction will have null or close to zero values, like R2 in the MUT state in Figure 5B.

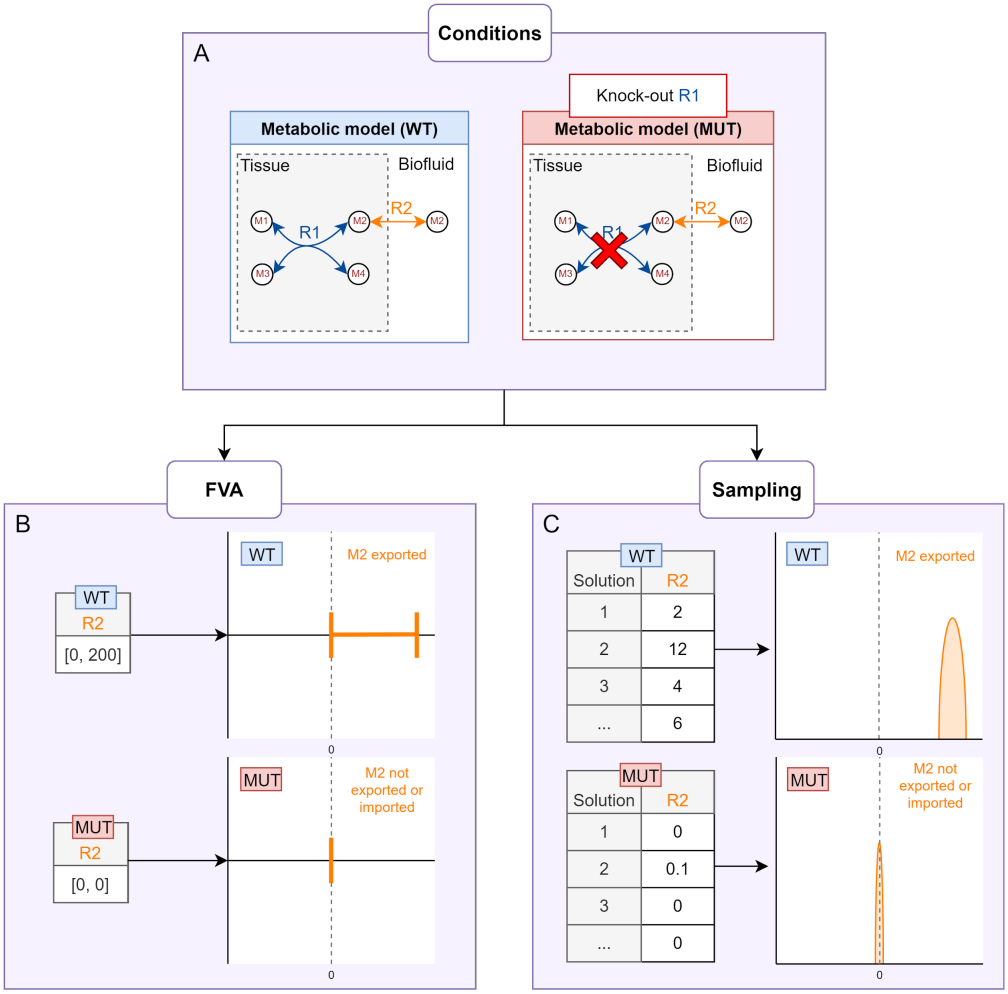


Figure 5: Flux Variability Analysis (FVA) and sampling for simulating fluxes in different conditions (A). The resulting flux values to be compared differ depending on the method used. FVA generates minimum and maximum possible flux values, shown as intervals (B), whereas sampling generates many values within those bounds, shown as distributions (C).

In [120], the authors developed a workflow for using condition-specific extracellular metabolomics data to predict intracellular fluxes. They created multiple versions of a subset of Recon2 to represent different conditions and sampled the fluxes of all reactions. To compare flux distributions between conditions, they compared median values and defined a minimal change of 10% to be considered as different. This mainly serves to evaluate shifts in flux distributions, and determine differences in metabolism between two leukemic conditions. This predicted a higher use of glycolysis for one model and a more oxidative phenotype for the other, supported by experimental validation, gene expression data, and the fact that leukemia cancer cells are known to rely more on glucose to support proliferation, while remaining heterogeneous.

A different method of comparing flux distributions is by calculating a z-score for each reaction. A z-score is a scoring metric which quantifies the number of standard deviations by which a value is above or below the mean of a distribution. In this case, a ”pseudo z-score” is calculated between two distributions using the mean and variance of that reaction in two conditions, as shown in [121, 122]. For this thesis, z-scores refer to this pseudo z-score applied to scoring the difference between two distributions. The z-score helps take into account the variation and uncertainty in the flux distributions, which is not possible using FVA or a comparison of medians, as well as the shifts in values. This can be taken further by calculating a p-value to evaluate the significance of the flux distribution change but can be difficult to interpret or choose a cutoff for.

3 State of the art in predicting biomarkers

Predicting biomarkers using metabolomics data is fundamental in discerning metabolic differences between different conditions, and between patients in differing cohorts or otherwise physiological states. From metabolomics data, potential biomarkers can be identified using classical statistical approaches

such as partial least squares discriminant analysis (PLS-DA), a supervised technique which maximises the variance between predefined classes [123, 124, 125]. This produces a list of certain features (in this case metabolites) that are able to separate these classes the most. The advantage of this compared with unsupervised methods like PCA is that sometimes PCA is not able to produce a meaningful separation between the expected groups, since by definition it maximises the global variance between individuals and not the part of the variance which is specifically associated to the biological factor of interest.

Metabolomics data can also be converted into experimental networks which can link the metabolite data based on different properties [126]. These can be based on mass differences between metabolites, structural similarities (using the raw data directly), and correlation data based on the metabolite abundances. Experimental networks can be used to annotate and identify features from untargeted metabolomics analyses, which can then be selected as potential biomarkers.

These biomarker prediction strategies are based on identifying biomarkers that already exist within the data. Alternative approaches exist which can be described as data-independent and are therefore able to select both measured and unmeasured features as potential biomarkers. Metaborank [127] is a network-based recommendation system inspired by how social networks recommend content to users. It uses experimental metabolic fingerprints to extend and recommend other metabolites of potential interest in relation to the initial profile, based on a GSMN. Knowledge-based networks can also be used to infer new information about potential biomarkers. For example, FORUM [91] harnesses information from the literature to provide new significant links between chemicals and biomedical concepts, by constructing a knowledge graph from published articles and annotated MeSH (Medical Subject Headings) terms. MeSH terms are part of the controlled vocabulary thesaurus used for indexing articles on PubMed. These links can help develop new hypotheses and

distinguish potential metabolites of interest in relation with a given condition.

3.1 Previous work on predicting biomarkers using CBM

The first paper to predict biomarkers using genome-scale metabolic networks and CBM was published by Shlomi *et al.* [128] in 2009. The study focused on using the human network Recon1 [81], along with known IEM diseases to predict 20 specific metabolite increases or decreases in the extracellular compartment (see Figure 18). 17 IEMs were modelled in the network as gene KOs using the known genetic mutations for each IEM from OMIM (Online Mendelian Inheritance in Man) [1], a database of human genetic disorders and associated traits. For each IEM, both a KO (knock-out) condition, where the reactions corresponding to the mutated gene are blocked, and a wild type (WT) condition, where the same reactions are forced to be active, were simulated.

The study used FVA to generate a pair of flux boundary values for each of the 20 exchange reactions in each condition (KO and WT), for each disease. These upper and lower bound values were then compared between WT and KO to predict an increase, decrease, or no change for that metabolite exchange. Therefore, for a given IEM, this results in a predicted change direction for a metabolite level in the extracellular space, which can be matched with observed data (of plasma levels) from OMIM.

This work was replicated by Mondeel *et al.* [129] in 2018 in order to improve the reproducibility of the project, and was published in the ReScience C journal. The original Shlomi *et al.* paper did not provide source code, and even if it had, it most likely would have been in Matlab, which is the standard for constraint-based modelling. Recently, Python has seen an increase in usage for CBM, with the release of the cobrapy package [130] in 2013. This replication was coded in Python and shared freely on github.

In a 2013 study, Thiele *et al.* [83] published the next human model in the

Recon lineage, Recon 2. In this paper, a similar analysis was carried out using FVA, this time on a wider range of 49 genetic diseases and 54 metabolites, using a gold standard created in [131]. This resulted in a larger table of predictions of metabolite increases and decreases, each compared with the corresponding observation in the gold standard (see Figure 19). The method was able to predict many relevant biomarkers, but the table was shown without false positives (prediction of a change where there is none) and negatives (absence of a prediction where a change was observed) in the paper. Their results are presented and discussed in Chapter III, and thanks to their published code, for this thesis, I reproduced this table from Thiele *et al.* to demonstrate the reproducibility of the method, as well as to highlight the missing false positives and negatives. This included all metabolite changes using their code, data and model, as well as versions using a different model, parameters, and an adapted version of their Matlab code in Python. This is also shown in the beginning of Chapter III.

3.2 Critical assessment of the use of CBM methods for biomarker prediction

FVA does well to evaluate the ranges of flux values that reactions in a metabolic network can take. By using the maximum and minimum bounds, we can gain information about the capacity of a reaction and how it can be increased or decreased. The size of the interval also indicates how adaptable a reaction is: a reaction with extremely tight flux boundaries cannot function in many different scenarios and can be considered as a non-flexible reaction. However, beyond this information, FVA does not provide information on the flux values within the boundaries. We could imagine any number of flux distribution configurations within these boundaries, such as uniform or bimodal distributions, but without evaluating this space, we cannot know which values are the most frequent and

we may make incorrect assumptions because of this. For instance, the mean of the minimum and maximum values (a.k.a. the halfway point) could be calculated but this relies on the assumption that the internal distribution is either uniform or normal and perfectly centred between the bounds.

FVA has another underlying limit when comparing bounds between conditions: there are multiple ways of defining the comparison between two intervals. This can especially become difficult when the intervals are both positive and negative (reversible reactions): for example, if a reaction has bounds of [-100, 100] in one condition and [-50, 300] in another, can this be considered an increase or a decrease of flux? The total flux interval has increased but the flux in one direction has decreased. This is one of many cases where flux intervals can be ambiguous in determining an increase or decrease.

Evaluating any sort of metabolite or biomarker prediction using traditional contingency tables, recall and precision is difficult due to the nature of metabolomics measurements and the available "truth" datasets. A metabolic model contains all known metabolites involved in metabolic reactions, but metabolomics methods are not able to detect and annotate all of them. Not every metabolite has been measured in patients of diseases, meaning a missing experimental value does not mean absence of change for that metabolite. This results in many cases where metabolites are predicted to be of interest while they are not detected by typical assays. In these cases, the predictions could be correct while being considered as a "false positive". New methods must be developed to take into account the nature of metabolomics data, and visualising the results of these analyses requires a better understanding of what we can and should show.

To predict biomarkers using GSMNs, the network must contain as much relevant knowledge of metabolism as possible. As genome-scale reconstructions have improved, models have increased in size, especially for well-studied organisms like the human metabolic network. Knowledge of new metabolites, reactions and their interactions with metabolic genes has increased, leading

to more connected networks. Globally, any calculation or simulation will take longer on a larger model and require more resources than a smaller one, especially for exponentially intensive calculations like random sampling. Sampling an entire network such as Human1 can take several hours on a multi-core computing cluster. This of course improves with faster CPUs and more memory availability, as well as better algorithms.

Initially, CBM methods were developed in MATLAB [132], a proprietary software platform for mathematical calculations, plots, and algorithms. An entire toolbox, COBRA [133, 134], was and is being developed for MATLAB and includes CBM algorithms and functions for many different methods. However, because MATLAB is proprietary and its code is in their dedicated programming language, the spread of information and learning is limited to the availability of these resources. COBRApy [130] is a Python equivalent of COBRA which has many of the same functions coded for use in Python.

4 Thesis objectives and outline

Metabolomics is a powerful approach for deciphering metabolic modulations through the identification of metabolic profiles. Nevertheless, metabolomics remains sparse in terms of metabolome coverage both for analytical and metabolite identification reasons. Experimental design can be complex when a compromise has to be made on how to measure metabolites the most precisely but also the most cost effectively. Having prior knowledge on metabolic endpoints of modulations could help target specific metabolites or entire compound classes.

Additionally, understanding the metabolic perturbation behind a phenotype is not always obvious as the effects of a disruption can propagate throughout metabolism and be unintuitive to the human eye.

During my PhD, the goal was to develop and improve upon the simulation of

metabolic profiles using genome-scale metabolic networks, and show how these predictions can aid in extending experimental metabolic profiles. We also wanted to help discern the links between a disruption (metabolic state) and what we can measure (metabolic profile).

This thesis is split into five parts: in Part II, I describe how I built a novel approach of using random sampling to predict metabolic profiles. I then provide applications of predicting metabolic profiles with concrete examples in Part III, followed by an exploratory approach of using simulated metabolic profiles for benchmarking pathway enrichment analyses in Part IV. Finally, the discussion, conclusion, and perspectives can be found in Part V.

Chapter II

Methodology: Developing a new approach for metabolic profile prediction

While previous modelling methods have been used to predict biomarkers, the goal of my thesis was to predict entire metabolic profiles by capturing, for each metabolite, its amplitude of variation between the control and the condition under study. This variability can be evaluated, and metabolites can be scored and ranked in order to prioritise the ones to be measured and annotated during metabolic profiling.

1 Main questions and design

The major question for this thesis was: can we go beyond predicting a few predetermined biomarkers for specific conditions, and simulate new scenarios as well as large panels of metabolites (metabolic profiles)?

Figure 6 shows the parallel between using metabolomics to produce experimental metabolic profiles (Figure 6A), and modelling metabolism *in silico* to simulate metabolic profiles (Figure 6B). The example profiles in this figure show the possibility of predicting new metabolite changes which weren't

detected experimentally (M1), and conversely, the inability for modelling to predict some metabolites (M6) (due to missing metabolites in the network for example) hence highlighting the strong potential to combine both experimental and *in silico* approaches.

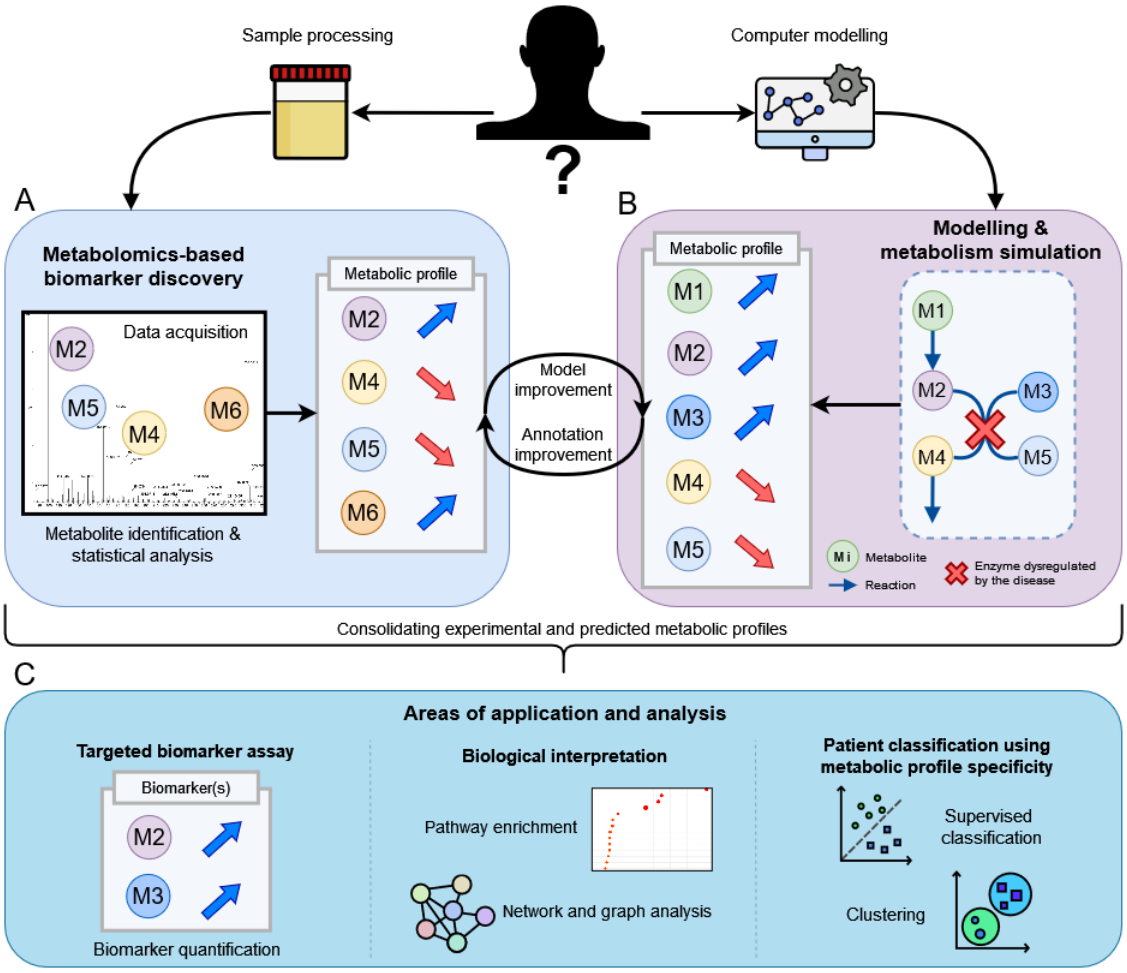


Figure 6: Combining metabolomics profiling with simulations of metabolism. A: Experimental-based biomarker discovery produces metabolic profiles containing detected and annotated metabolites along with a concentration or fold-change value. B: Metabolic disruptions can be modelled to simulate metabolic profiles similar to those generated using metabolomics. C: By combining information from both types of metabolic profiles and improving both experimental annotation and in silico models, various approaches can be used to improve our knowledge of given biomarker sets, affected pathways, and patient disease classes.

Computational solutions can be used to fill the gaps in experimental observations by providing a recommendation list of metabolites which are expected to be altered in the studied condition. This list can be used at different steps of the metabolomics process (Figure 6C). Firstly, it can be used upstream

of the experiment to select "the most suitable" analytical platform and set-up (e.g. if mostly lipids are expected to be affected, a lipidomics setup will be favoured). The benefit of predicting profiles is also downstream of the analysis for annotation purposes. Raw data can be mined to look directly at the predicted metabolites which will accelerate the process of identification. This prediction can be used to select the right set of standards to be analysed to reach level 1 annotation. Finally, they can be of added value to fill gaps in biochemical interpretation by suggesting metabolites (and related pathways) which could be of interest for the biological comprehension of a disease.

To improve upon previous studies in the prediction of biomarkers using FVA, a different CBM method was selected. We chose a random sampling approach to sample all exchange reactions in the network to get a finer grain of detail in the prediction results. This sampling method combined with scoring and ranking the exchange reaction flux differences composes the novel methodology described in this chapter.

2 Overview of the methodology behind predicting metabolic profiles and its associated methods

2.1 Metabolic profile simulation methodology

By combining the flux simulation of exchange reactions with a network disruption, different flux values can be obtained for both a healthy (default network) and a disrupted condition. Each metabolite's exchange reaction fluxes can then be compared in order to determine an increase, decrease or no change between the two conditions. As an example, if a metabolite's exchange reaction results in a high export in the healthy condition, and a flux of zero in the

disrupted condition, we can then say that this disruption is expected to cause a decrease of that metabolite in the biofluid compartment. By simulating this for every exchange reaction in the network, we can produce an *in silico* metabolic profile associated with a given disruption.

This CBM methodology can be used to predict which metabolites will be more or less released in biofluids by using an organism-specific metabolic network in conjunction with a metabolic disruption. Indeed, in metabolic networks, some metabolites can be transported in and out from the internal compartment (cell or tissue) to the external compartment (e.g. biofluid or cell culture medium) usually using a single specific exchange reaction. For the *in silico* prediction of biomarkers these exchange reactions can be used to model the in/out flux of metabolites between tissues and circulating biofluids like blood or urine. This is why, in the context of metabolic profile prediction, the focus must be on these specific exchange reactions from the metabolic network in order to predict the equivalent of "biofluid metabolite level changes" using CBM. A break-down of this methodology is shown in Figure 7, using a simple metabolic network to compare flux simulations in healthy and disease conditions, and resulting in a ranked list of metabolites which change the most between the two conditions.

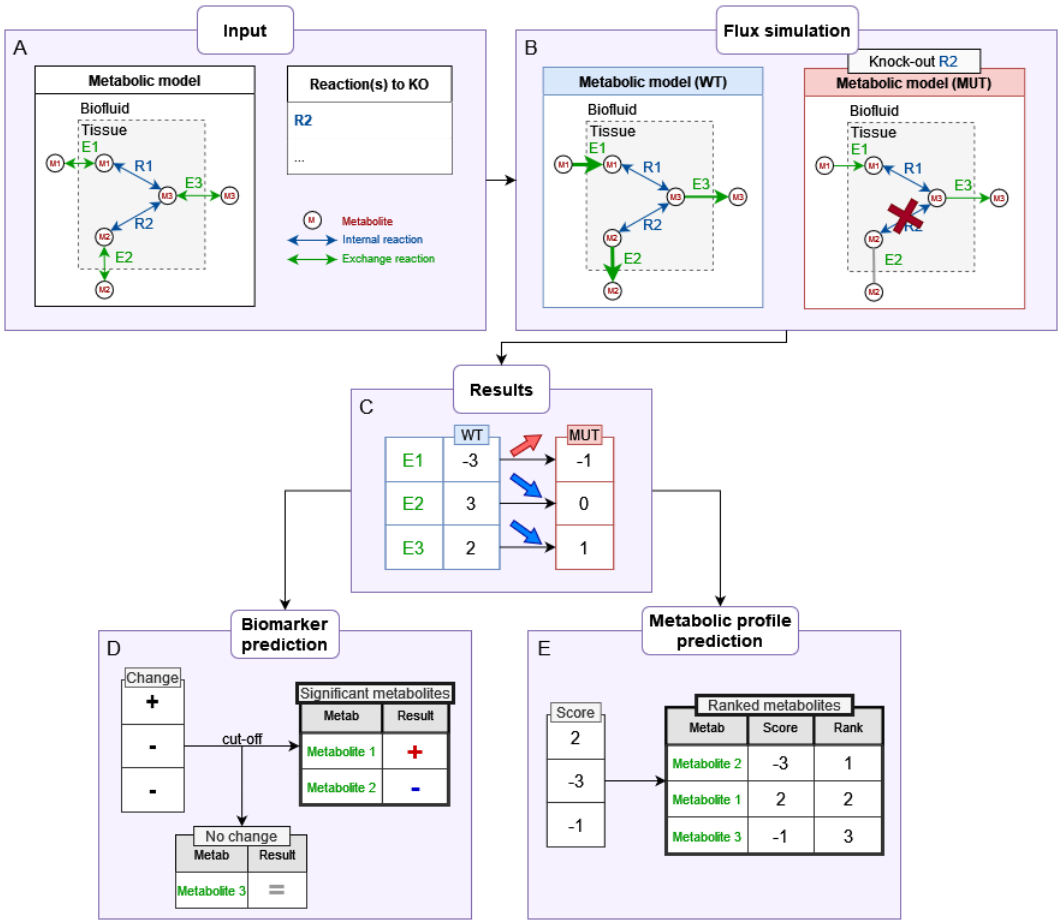


Figure 7: Methodology for the comparison of flux values and prediction of metabolite ranks using a simple network (A) in two conditions (B), with single flux values (C). The methodology from Shlomi et al. is shown in (D): each exported metabolite will have an associated change score for a given pair of conditions. (E) shows our methodology of scoring and ranking by absolute value among all of the metabolites in the network.

Metabolic network models are usually undetermined, i.e., there are not enough constraints in the model to determine a unique solution for the mass balance system of linear equations [104]. For example, in the wild type (WT) (Figure 7B), any other value for E1 would still ensure the steady state, as long as the other reactions' flux values are changed accordingly to compensate. This is why it is difficult to define any one exact value for a reaction, and it is generally necessary to evaluate the range of possible flux values for all reactions through various CBM methods.

2.2 Simulating metabolic conditions

Setting up the metabolic state of the GSMN is the first main step of simulating a metabolic profile for a given condition. First, the model must be chosen. This choice can be more or less complex depending on the available models for the organism of interest. Some organisms only have one or two existing metabolic models, whereas others have many different versions, which can contain new informational updates.

GSMNs can also be contextualised to a specific cell type or tissue from a global version of an organism's network, by integrating transcriptomics or other biological data [135, 136]. This creates context-specific models which can be compared directly or used to simulate fluxes, which can then be compared. Generally, these models differ from the base network on a relatively large scale: several tens to hundreds of genes could be modulated to simulate one condition.

In this thesis, the focus is more on smaller-scale perturbations, simulating a metabolic disruption due to a genetic mutation rather than the genetic expression of entire cell type or tissue. Simulating these perturbations requires a list of genes or reactions to be disrupted as input. The genes or reactions have to be mapped to the chosen model, which can be a time-consuming manual step if the model does not contain external IDs, or if the reactions or genes of interest are not all present in the model. Additionally, if using genes as input, the information contained within the network is used to extract the reactions linked to these genes via GPR relationships, which are not always correctly defined in the model and this often requires manual curation to check the validity of the GPR links.

2.2.1 Knock-out

The knock-out (KO) state is the easiest state to simulate once the reactions of interest have been identified in the model. It consists of simply setting all reactions to KO as blocked reactions, i.e. setting both of their new flux bounds to

$[0, 0]$, replacing the previous bounds:

$$ub_{new} = lb_{new} = 0 \tag{4}$$

for a reaction R, with ub as upper bound and lb as lower bound.

2.2.2 Wild type

The wild type state may appear to simply be the default model, but in order to ensure a fair comparison with the disease state, extra parameters must be set in the model. To begin with, the WT state is created using the default network parameters (reaction bounds, biomass coefficients etc.). The most obvious approach would be to only use these default parameters to create this WT state. However, when simulating the fluxes of a reaction using the default flux bounds, there is a risk that the fluxes end up being too similar to those of that reaction in the KO state, since reaction bounds can be [0, 1000] for example, which means 0 is a possible flux value for that reaction. To avoid comparing two states where the fluxes for the reactions of interest are both zero or near zero, the reaction(s) to be knocked out in the KO state are forced to carry a non-zero flux in the WT state. This WT method is the one used in Shlomi *et al.* and Thiele *et al.* for their IEM biomarker predictions.

The maximal possible flux through the reaction to KO is determined by optimising for this reaction by setting it as the model's objective function to maximise. Then, in the WT model, the minimum bound is set to 5% of the maximum flux value (R_{max}) obtained from the previous maximisation. Forcing a minimum flux in the WT is why each WT is specific to a mutant (MUT) state.

$$lb_{new} = 0.05 * R_{max} \tag{5}$$

for a reaction R, with lb as lower bound.

2.2.3 Knock-down

In addition to completely knocking out reaction fluxes, they can be instead partially reduced using a new method developed for this thesis project. The reaction knock-down (KD) method is the following: instead of completely blocking one or several reaction(s), the MUT state can consist of a reduction percentage of the maximum possible fluxes for a given set of reactions. This is done by first calculating the maximum possible flux range using FVA in the chosen metabolic condition. Then, depending on if the FVA upper bound ub and lower bound lb are forward, backward or reversible, the flux reduction is carried out differently:

$$ub_{new} = lb_{FVA} + (r * (ub_{FVA} - lb_{FVA})) \text{ for forward reactions } ([0, 1000])$$

$$lb_{new} = ub_{FVA} + (r * (lb_{FVA} - ub_{FVA})) \text{ for backward reactions } ([-1000, 0])$$

$$\begin{cases} ub_{new} = ub_{FVA} * r \\ lb_{new} = lb_{FVA} * r \end{cases} \text{ for reversible reactions } ([-1000, 1000]) \quad (6)$$

By using this reduction strategy, each total interval space is reduced to $r*$ total range. For example, an interval of $[-1000, 1000]$ has a total range of 2000, therefore to reduce the range to 30% of 2000 (600), both bounds must be multiplied by 0.3, leaving us with $[-300, 300]$ and a total range of 600.

Figure 8 shows an example of different KD values on a reaction, from 10% flux to 100% flux. The reaction shown is therefore the perturbed condition (input of the simulation), and not an exchange reaction (output of the simulation). This highlights the approach consisting of reducing the maximum flux bound, as opposed to also increasing the minimum flux bound, which would not be coherent since it would imply forcing a higher minimum amount of flux (which could instead be used to simulate an increased enzymatic activity).

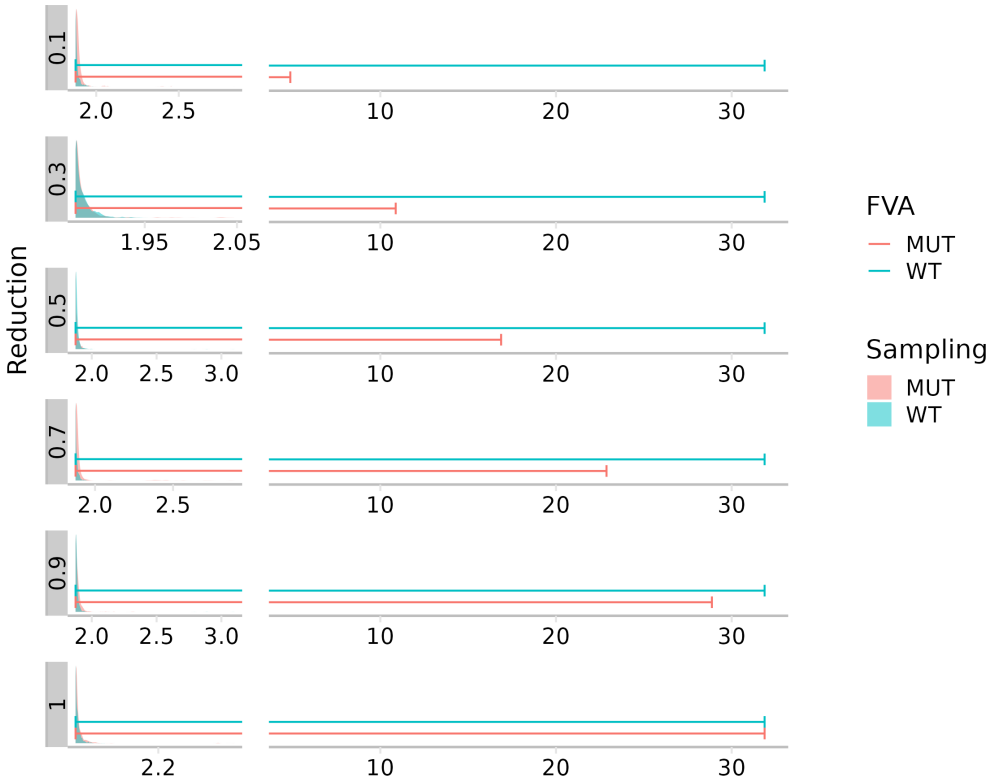


Figure 8: Different knock-down percentages on an example reaction, showing FVA bounds and the corresponding sampling distributions. 0.1 reduction corresponds to reducing the upper flux bound to 10% of its maxmimum value.

The result of reducing fluxes instead of fully knocking them out is that exchange reactions are affected to a lesser extent. The two following examples (Figure 9 and Figure 10) show two different levels of disruption due to KDs.

In the MUT 0% flux state, both exchange reaction A and B are affected in the most "extreme" manner possible, shown by the FVA bound differences and the sampling distributions shifts in the top plots of Figure 9 and Figure 10. However, as the flux reduction percentage increases (Figure 8), meaning less extreme KDs, the changes are less drastic: in both cases, the distributions do not shift for any reduction value. For exchange reaction A, the FVA bounds are reduced to a lesser extent (compared with the 0% state) for 10% and 30% flux reduction. Exchange reaction B is not affected regardless of the reduction value.

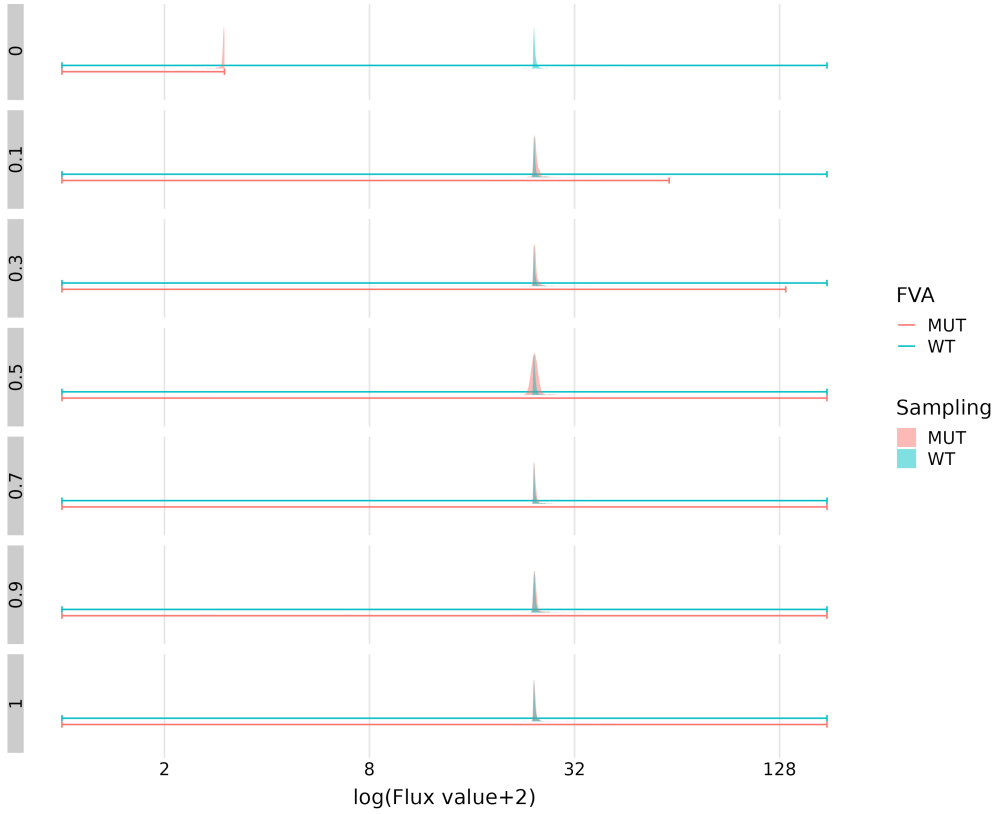


Figure 9: Flux bounds and distributions for **exchange reaction A** for different flux range knock-down values (y grid axis), which can be seen as knock-down percentages. The top plot row (0) corresponds to the full KO state. A knock-down of 0.1 corresponds to the MUT state reactions only having 10% of their maximum flux range. Flux values are shown on a log2(value+2) scale.

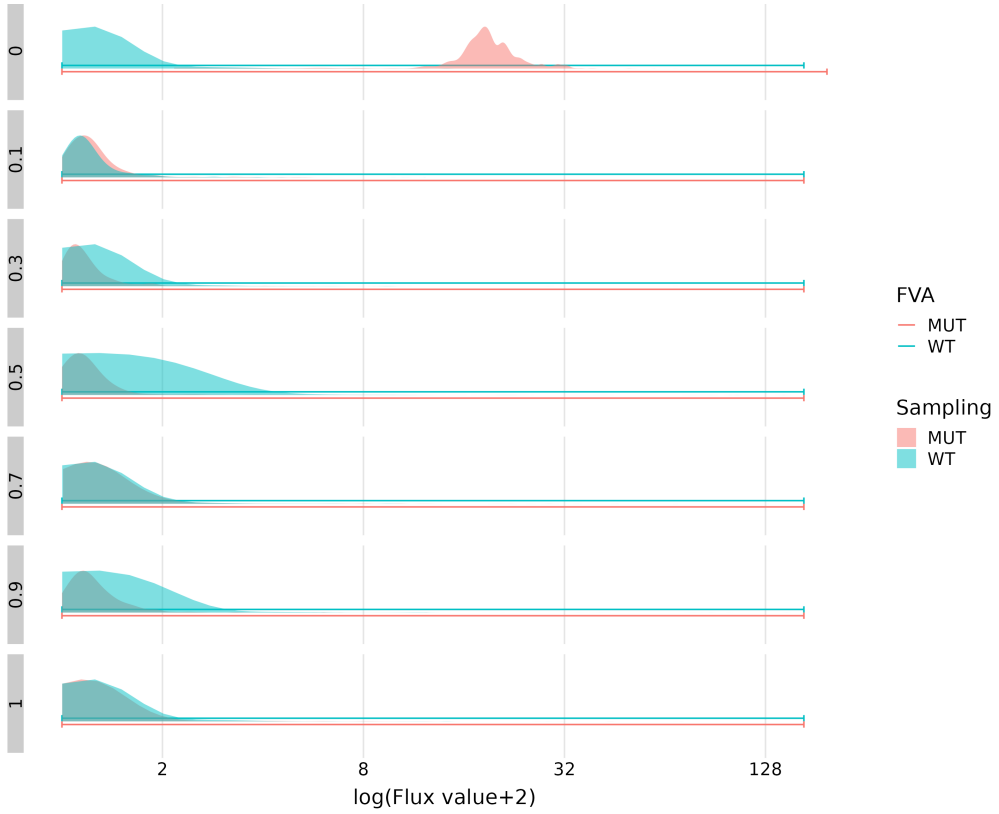


Figure 10: Flux bounds and distributions for **exchange reaction B** for different flux range knock-down values (y grid axis), which can be seen as knock-down percentages. The top plot row (0) corresponds to the full KO state. A knock-down of 0.1 corresponds to the MUT state reactions only having 10% of their maximum flux range. Flux values are shown on a log2(value+2) scale.

2.3 Random sampling methods

By exploring the many combinations of flux values that satisfy the model constraints, sampling provides an overview of the most frequently valid flux values for every reaction in the network. Indeed, any one possible solution is a specific combination of flux values for each reaction, and some flux values are more frequent than others, meaning that they appear more often in different valid solutions. These many solutions can therefore outline each reaction's flux distribution and thus reveal the most frequent fluxes along with the variability of the values, as shown in Figure 5. While sampling fluxes in GSMNs in general is not new, its application to the prediction of biomarkers and metabolic profiles is a novel approach developed during this thesis.

As described previously in relation to other CBM methods, sampling explores other possible (but not necessarily optimal as in FBA, or extreme as in FVA) solutions contained within the solution space. The solution space can be visualised as a high dimensional cone, as shown in Figure 11.

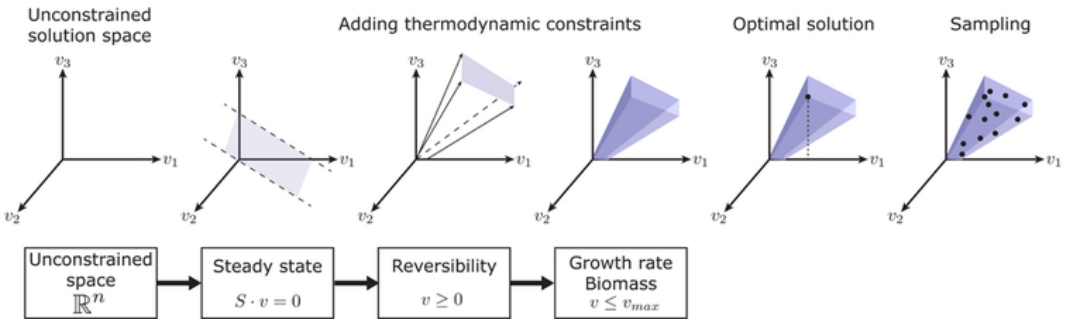


Figure 11: Constraint-based modelling solution space cones. From [133].

Sampling is an unbiased method of obtaining solutions as it does not involve selecting one single solution among many. Random sampling flux distributions not only provides a more detailed description of each reaction's flux behaviour but can also be represented visually as distribution density plots for a more intuitive comprehension of the fluxes passing through a reaction.

The flux solution space of metabolic networks is a convex polytope, usually in an irregular shape which is elongated or narrow depending on reaction constraints. This means that many samples are close to the edges of the solution space. When exploring areas close to the edges using a hit-and-run algorithm [137], each new sample is chosen using the previous sample and the direction is chosen from all possible directions. Being close to the edge means a significant portion of possible directions is unavailable which leads to closer samples.

Artificial Centering Hit-and-Run (ACHR) algorithms are better at exploring these irregularly shaped solution spaces. They do this by pushing for samples in elongated directions resulting in samples further away from each other. By using an estimation of the center of the solution space at each step and by generating warm-up samples before the main sampling phase, ACHR can explore these

specifically shaped spaces more thoroughly. ACHR is used by gpSampler [138], a popular sampler for exploring the solution spaces in metabolic networks.

The sampling algorithm selected for this project is optGpSampler [139], which improves upon previous metabolic network sampling algorithms. It is based on a ACHR algorithm, combining its warm-up phase with an optimised version of gpSampler's warm-up phase, as well as an optimisation of how sampling chains are managed. It has a Python interface and is implemented in cobrapy.

2.4 Scoring metabolite changes

2.4.1 Z-scores

In order to conclude on a change between two conditions, we need a method of not only comparing but also scoring the differences. Comparing two distributions is more difficult than comparing two values due to the multiple inherent properties of a distribution. This necessitates a different approach to improve the interpretation of distribution changes. A score is calculated for each exchange reaction in the model by comparing the samples between both states. We propose the use of a z-score to evaluate the shift in distributions weighted by their variance, based on the z-score used in Mo et al. [121].

Z-scores are calculated for each metabolite's exchange reaction ex_i between the WT and MUT. First, we sample a number (by default the total number of samples) of random pairs of values from the WT and MUT distributions. The collections of all MUT samples and WT samples for metabolite i's exchange reaction are MUT_i and WT_i respectively. A "difference distribution" dd_i is calculated by subtracting random pairs of values from both MUT_i and WT_i (Equation (7)). These random samples are not matched: the two WT and MUT values are not necessarily from the same sample step. The final z-score z_i is calculated by dividing the mean μdd_i by the standard deviation σdd_i of dd_i.

Equation (7): For the ith exchange reaction:

$$z_i = \frac{\mu(MUT_i - WT_i)}{\sigma(MUT_i - WT_i)} = \frac{\mu_d d_i}{\sigma_d d_i} \tag{7}$$

The z-score z_i is directional: a negative z-score indicates a decreased shift in the flux distributions from WT to MUT, and a positive z-score indicates an increased shift. A z-score close to 0 means that there is little difference between the distributions in the WT and MUT conditions. A z-score therefore represents the intensity and direction of one metabolite's shift in a specific condition.

Z-scores of all the exchange reactions in the network are used as a basis to rank all exchanged metabolites based on the intensity of the changes. Z-scores can also be used as-is or used with a threshold. Since both increased and decreased metabolites are of potential interest, this ranking is based on the absolute values of the z-scores (Figure 7E). This reveals the metabolites whose import/export behaviour changes the most between the WT and MUT, relative to every other exchange metabolite in the network.

Furthermore, ranking the z-scores by absolute value provides insight via the comparison of the list of the top ranked metabolites between different scenarios, as ranks are relative to the entire list of exchange metabolites and not quantitative and therefore do not require normalisation.

2.4.2 Exploring alternative scoring methods

While a z-score was chosen for this metabolic profile prediction approach, other scoring methods were analysed and compared to make sure it was the best suited method, such as the subtraction of distribution means. In the following figures, the example from Chapter 3 Section 3 was used for illustrative purposes. The example is from an mGWAS cohort which analyses SNPs significantly associated with various experimentally measured metabolites. When modelling one example SNP, the simulation predicts the 1497 exchanged metabolites from

the network and compares them to the list of 20 experimentally significant metabolites (shown in red in the following figures). The list of 1497 metabolites is ranked based on the z-score as well as other methods, and them compared between each scoring method.

Figure 12 shows the comparison of using a z-score to compare distributions, vs subtracting the means of the distributions (left panel) or medians of the distributions (right panel). For each pair of metrics, the top ten metabolites for each given metric are displayed with their labels (red being those in the experimental signature) and with dots located on the y-axis. The y-axis corresponds to the position in the ranking of these metabolites among the 1497 metabolites using the corresponding metric (the top being the first metabolite while the bottom is the last metabolite for the given ranking).

Using the z-score as a ranking metric predicts more experimentally observed metabolites in the top 10 than when using the difference of the means or medians to rank metabolites. Additionally, two experimentally observed metabolites, oleate and palmitate, that were not in the z-score top 10 are highly ranked when using the mean- or median-based rank. However, most of the other top 10 ranked metabolites using the mean differences appear to be unrelated and non-specific (H+, CO2) or vague (metabolite pools) in modelling terms. The metabolites in the top 10 of the z-score-based ranking are ranked in the top $\sim$200 when using the difference between means or medians to rank metabolites, which shows that while they do not predict as well as the z-score for these metabolites, they are relatively well-ranked.

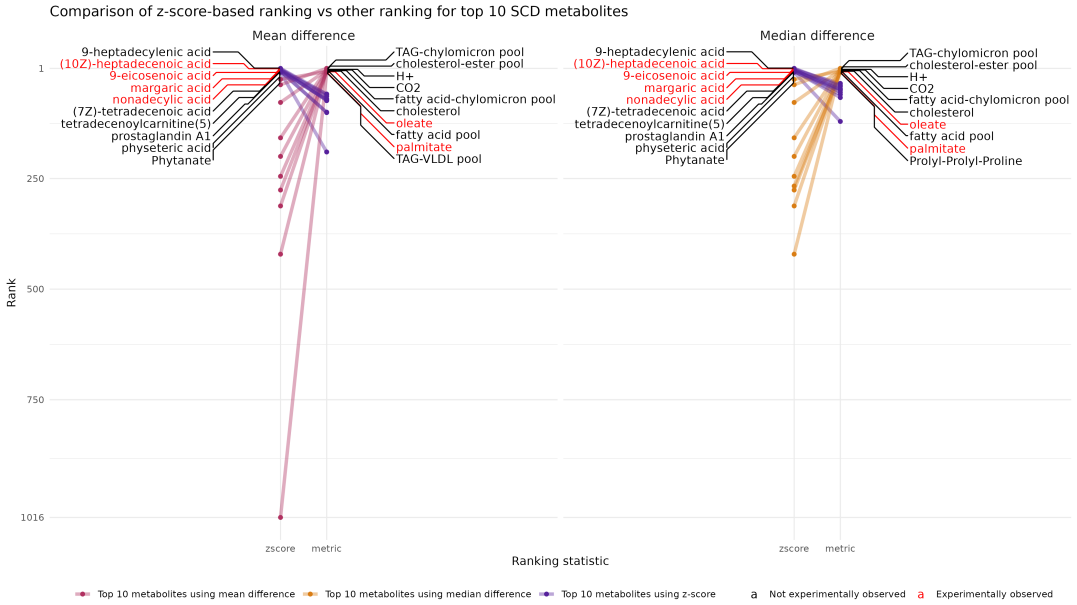


Figure 12: Z-score based ranks (left side of each plot) vs using the difference between means (right side of left plot) or medians (right side of right plot) to rank metabolites. The metabolite labels are all ranked in the top 10 by each method. The metabolites highlighted in red are differentially abundant in the example dataset, whereas those in black are the rest of the top 10 metabolites which are not experimentally abundant. For example, both TAG-chylomicron pool and oleate are ranked in the top 10 when using the mean difference as a ranking metric instead of the z-score (right side of the left plot), and TAG-chylomicron pool is in black because it is not observed to be experimentally abundant, whereas oleate (in red) is significantly abundant in the example experimental data.

Other metrics were compared to z-score ranks: the ratio of means, medians, 1st quartile and 3rd quartile were computed and used to rank the metabolite changes (Figure 13). In these cases, the ratio metrics were not adapted as they did not contain any of the experimentally abundant metabolites in their top 10, and the top 10 z-score predictions were poorly ranked, regardless of the metric used to compare with the z-score ranks.

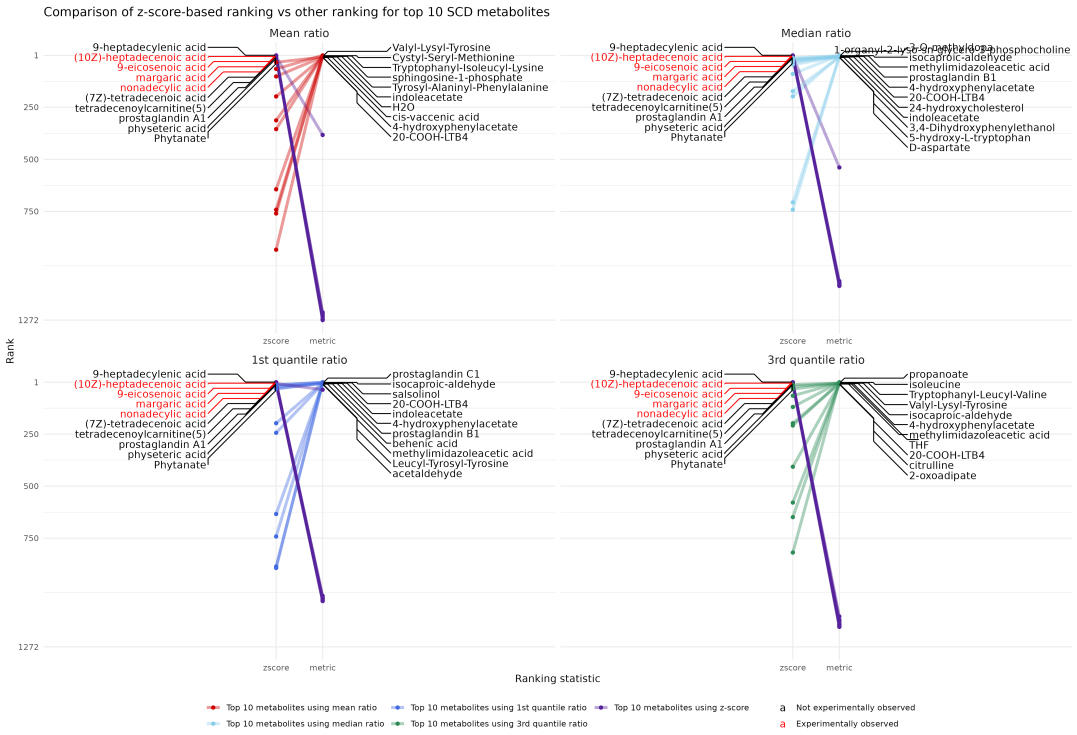


Figure 13: Z-score vs other metric-based rankings. In each subplot, the z-score rank is on the left with the same total order of metabolites. On the right of each subplot is a different ranking generated using a different scoring metric, such as the ratio between the means of the two distributions for each metabolite. The metabolites highlighted in red are those experimentally observed in the example study.

When comparing differences (Figure 12) and ratios (Figure 13), the best metrics out of the two appear to be when subtracting means or medians rather than a ratio, which is akin to a fold change. Indeed, using fold changes in this scenario can be limiting due to its bias towards values close to zero. Both the mean and median differences result in a similar top 10 and rank the z-score top 10 metabolites below the 200 mark.

Ultimately, the cons of using the other metrics studied here outweighed the pros, since the ranks improved marginally while the top predicted metabolites were flooded with metabolite pools and irrelevant metabolites. For the rest of this thesis, we used the z-score to rank metabolite changes as it provided the best results when looking at the top of the ranked list. Indeed, using the bottom half of the list can be complex due to the similarity in z-scores for the low-ranking metabolites (see Figure 31).

2.5 Final pipeline: summary through a toy example

Let us focus this methodology on a toy example consisting of the network in Figure 14 for two metabolic disruption examples. Example 1 is centred on reaction $REF1$ (shown as a blue square) which takes metabolites E and X as substrates and produces metabolites F and Y (shown as blue circles, with F in purple due to it being shared with the second example). Example 2 is directed at the reactions involving metabolite C, B and F (red circles, with F in purple): RBC and RCF (shown as red squares). Most metabolites in the network have an exchange reaction which imports or exports them between the cell and the biofluid compartment.

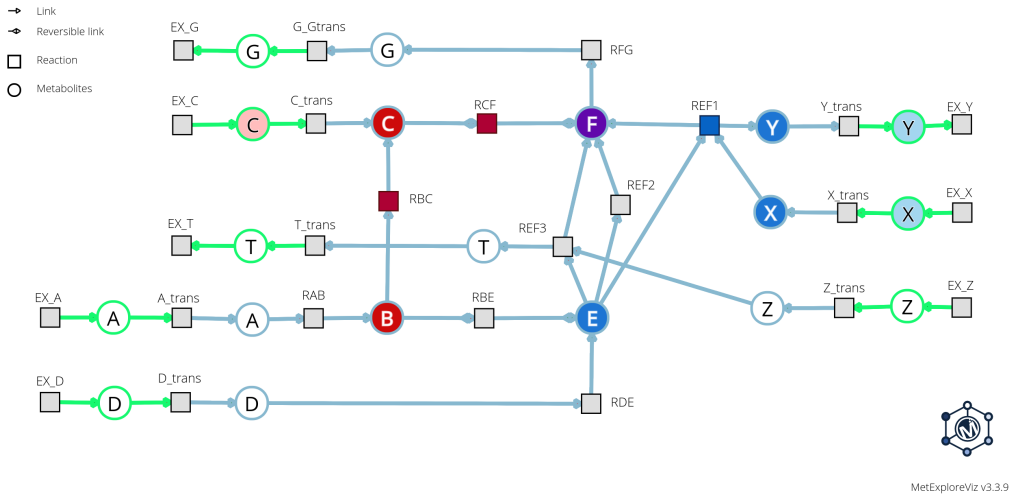


Figure 14: Toy network composed of metabolites (circles) and reactions (squares), visualised with MetExploreViz. Blue links are the cytosol compartment. Green links are the extracellular compartment. The blue square is the knocked out reaction for Example 1, with its substrates and products shown as blue circles. The light blue circles are the most disrupted metabolites in Example 1's flux simulation condition. The red squares are the knocked out reactions for Example 2, with their substrates and products shown as blue circles. The pale red circles are the most disrupted metabolites in Example 2's flux simulation condition. Metabolite F is shown in purple due to it being shared between both examples.

The goal is to predict metabolite exchanges with the biofluid compartment between a healthy condition and a disease condition. In this example, two separate conditions are simulated: a total knock-out of reaction $REF1$, and a total knock-out of reactions RBC and RCF. By following the pipeline defined

previously, the following steps are carried out:

1. Choose the model: in this case, the model is a simple toy model.
2. Set up the WT state by forcing a minimum amount of flux through reaction $REF1$ in Example 1, and through RBC and RCF for Example 2.
3. Set up the MUT state by setting reaction $REF1$'s bounds to 0 in Example 1, and RBC and RCF's bounds to 0 in Example 2.
4. Sample the fluxes of all exchange reactions in both WT and MUT states independently, for both examples.
5. Compare the two flux distributions for exchange reaction by calculating a z-score for each.
6. Rank the z-scores to gain a list of the most changed metabolites between the two conditions.

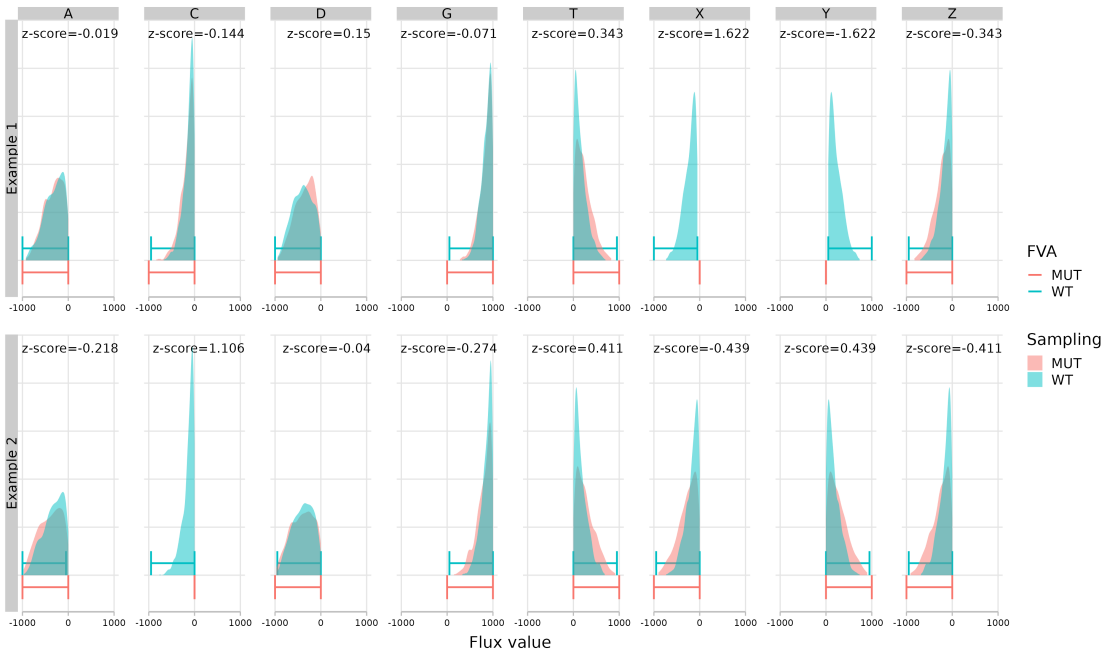


Figure 15: Toy flux sampling and FVA results for two example knock-out conditions. Blue and red represent the flux distributions (sampling) and flux bounds (FVA) for the WT state and MUT state respectively. The top row is Example 1, consisting of the knock-out of $REF1$, while Example 2 (bottom row) is the double knock-out of both RBC and RCF.

Example 1: in the WT state, flux can go through reaction $REF1$ and produce metabolites F and Y. Y is then exported into the biofluid compartment, while X is imported, as shown by the blue distributions in Figure 15. By blocking $REF1$ in the MUT state (red), Y can no longer be produced since it was only able to be produced by this one reaction. Its export reaction fluxes are therefore at 0. Similarly, X can no longer be used since only $REF1$ used X as a substrate, so its export reaction EX_X's fluxes are also at 0. There is therefore less Y in the biofluid in the MUT compared to the WT since it is no longer being exported, and more X in the biofluid in the MUT since it is no longer being imported. More subtly, metabolites T and Z have slightly increased export and import fluxes respectively in the MUT state due to repercussions in the network, as shown by the z-scores. T and Z are both involved in $REF3$ which is linked to F, a substrate of the blocked reaction $REF1$. The metabolites can be ranked by their z-scores to form a list of most changed metabolites (Table 1), with X and Y at the top, followed by T and Z, then G, D, C, and A.

Example 2: C is the metabolite the most affected by the KO of RBC and RCF: in the MUT state, its exchange reaction flux is 0. The only way C can be depleted is through RCF, and it can only be produced by RBC, meaning if they are both blocked, no flux can be carried through EX_C.

Rank	Example 1	Example 2
1	X / Y	C
2	T / Z	T / Z
3	G	X / Y
4	D	A
5	C	G
6	A	D

Table 1: Ranked lists of the metabolite predictions for both toy example conditions. Ties are shown in the same cells with "/".

3 SAMBA

Once the methodology was developed, I put everything together into one pipeline for ease of use, tracking past runs, versioning and improving development. The toolkit is called SAMpling Biomarker Analysis (SAMBA) and consists of two main parts: a Snakemake Python-based workflow for running the metabolic flux simulation (left side of Figure 16), known as SAMBAflux, and an R library and Shiny app for visualising and ranking metabolites (right side of Figure 16), known as SAMBAR. The steps described in the previous section are shown in Figure 16, and go from model condition set-up, through flux simulation, all the way to extracting a predicted ranked list of most changed metabolites.

The code for the SAMBA project is freely available at:

- https://forgemia.inra.fr/metexplore/cbm/samba-project
- https://doi.org/10.5281/zenodo.8369624

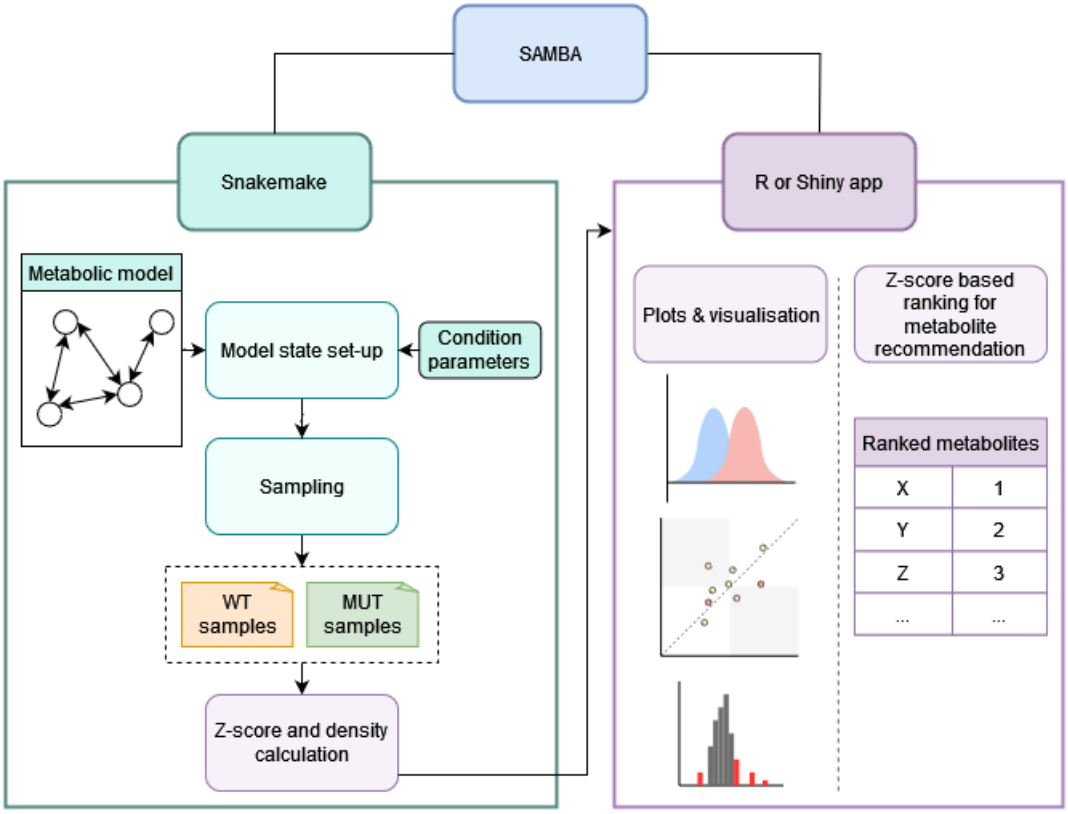


Figure 16: The SAMBA toolkit. SAMBA is split into two parts: SAMBAflux for the metabolic network flux simulation, and SAMBAR, which can read in the results from SAMBAflux, rank metabolites, and create plots.

3.1 SAMBAflux pipeline

Once SAMBA is installed, the user can provide a genome-scale metabolic network for a given species, tissue or metabolic condition, along with the desired metabolic perturbation(s) in the form of gene or reaction IDs. Multiple independent conditions can be run using one input file, as the Snakemake pipeline handles parallel jobs automatically. It can be launched locally, for small analyses, or on a high performance computing cluster for larger models and running multiple jobs in parallel.

Under the hood, SAMBAflux uses the provided model to prepare two model states: a WT state and a MUT state. It then generates flux samples for all exchange reactions (by default). SAMBAflux then calculates the difference between each pair of distributions and returns a z-score for each exchanged metabolite.

Various parameters can be modified by the user in the configuration file. Network modelling parameters such as biomass optimisation, initial exchange reaction bounds, and the reactions to knockout all affect the "biological" side of the simulation and the model's two states. The number of samples, thinning factor, the solver to use and the number of processors influence the sampling efficiency. Finally, the user can specify which reactions to output and where to output them to. In order to predict metabolic profiles, by default only the resulting flux values for the exchange reactions are exported. This results in two (WT and MUT) tabular files with rows as samples and columns as exchange reactions, with the exchange reaction names as column headers (see Table 2).

Sample	EX_A	EX_B	EX_C	...	EX_m
1	231	3	24	...	704
2	225	43	25	...	899
3	503	27	22	...	835
...	...	...	...	...	...
n	173	37	22	...	803

(a) WT

Sample	EX_A	EX_B	EX_C	...	EX_m
1	0	5	26	...	332
2	0	15	27	...	302
3	0	4	20	...	297
...	...	...	...	...	...
n	0	16	23	...	354

(b) MUT

Table 2: Example output sampling file format for a WT state (a) and a MUT state (b), for n samples and m exchange reactions.

Random sampling is done using Python code written for SAMBA, based on the cobrapy [130] Python package. The code uses the CPLEX 12.10 solver by default and uses the optGpsampler algorithm [139] to sample from the reaction flux solution space. optGpsampler begins with a warm-up phase to select starting points (by running a preliminary FVA on each reaction), followed by uniform sampling within this feasible solution space. Because each sample is selected

from the solution space directly, there is no sample rejection since this would be extremely inefficient to do on genome-scale models. A thinning parameter of k (default k = 100) means that every k sample is saved and the rest is discarded in order to reduce intersample correlation. For large models such as Recon2 and Human1, 100 000 samples with a thinning of 100 were used. Convergence tests are shown in Section 4.

Sampling can be run on a local computer for smaller models, but it needs a certain amount of resources to run correctly. More specifically, the amount of RAM required increases with the size of the model, and more CPUs will help generate the samples faster.

For this thesis project, the larger metabolic models (Recon2 and Human1) were sampled using a computer cluster which uses 16 cores and 128GB of RAM for each job. The cluster we used is the Genotoul computational cluster which has about 3000 cores / 600 threads, 36 Tera Byte memory (3TB on a SMP machine), Infiniband interconnection (QDR/FDR), parallel file system (GPFS). For one condition, sampling using Recon2 takes approximately 30 minutes, whereas sampling Human1's fluxes takes around 2 hours, both with 16 cores.

3.2 SAMBAR and RShiny

SAMBA outputs a compiled dataframe of z-scores in the form of a .tsv file, along with an density file (.json file) containing density approximations for each sampling distribution, for each condition. This data can then be taken into R to be analysed and plotted using SAMBAR functions, or similarly used in the SAMBA Shiny web app.

The Shiny app is currently online at https://samba.sk8.inrae.fr/ (see Figure 17), and the code is freely available for running the Shiny app locally through R. The app provides a user interface for reading in sampling density files and z-score files, and plots various representations such as distributions

and scatterplots for selected conditions. The z-score can have a threshold applied to it in order to filter the sampling distribution plots to the top N most changed metabolites. The user can vary the threshold and the plots will update dynamically.

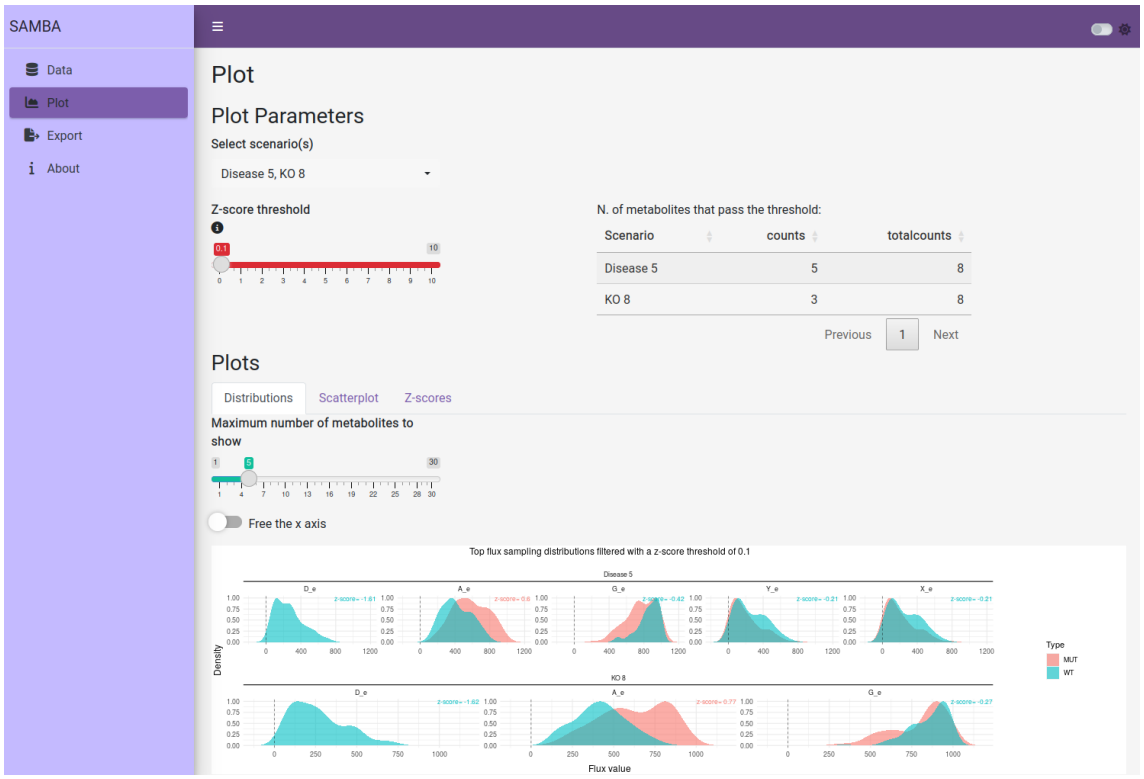


Figure 17: Screenshot of SAMBAR Shiny on a toy network, for two test conditions. The top most deregulated metabolites are shown for each condition simultaneously as rows in the sampling distribution plot.

Chapter III

Results Part I: Prediction of key metabolites using metabolic flux simulation

By exploiting experimental results with computational tools, new insights can be gained to improve both the prior knowledge required to design experiments as well as the interpretation and analysis of data. In terms of biomarker screening, simulations are able to predict certain metabolic markers of genetic diseases, known to be used in diagnosis, using constraint-based modelling and genome-scale metabolic networks. This chapter will go over my reproduction of this work, followed by comparison to the new approach presented previously as well as its applications the prediction of entire metabolic profiles, as opposed to a few biomarkers.

1 Reproduction of Thiele *et al.* results

1.1 Previous work by Shlomi *et al.* and Thiele *et al.*

In 2009, Shlomi *et al.* [128] used FVA combined with a human GSMN, Recon1, to predict 19 biomarker changes for 17 IEMs. The methodology they used was

the exchange reaction-based approach described in Chapter II, which produces, for each metabolite's exchange reaction, a lower and upper flux bound in the WT and MUT states. These bounds are the most extreme flux values in both directions that a reaction can take. The two intervals can then be compared between both states and if there is a difference between them, the metabolite is described by their method as increased or decreased in the biofluid compartment.

The figure they published in their paper can be found below (Figure 18), which describes the predicted flux increases and decreases (blue and red coloured cells respectively) for the aforementioned metabolites, compared with observed changes in patients of each IEM ('+' and '-').

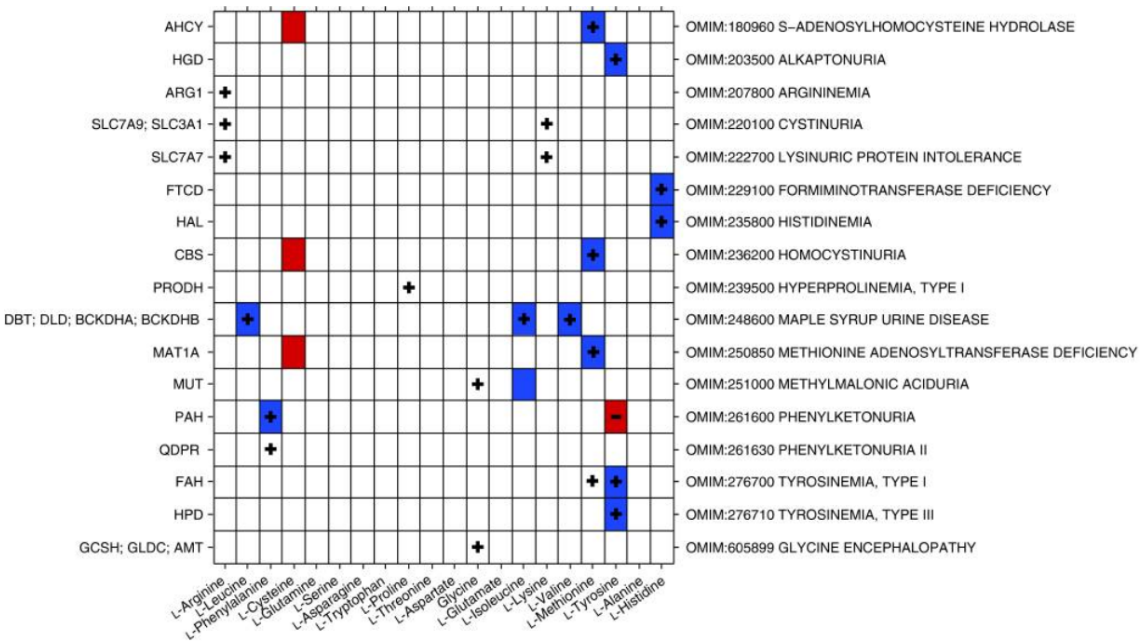


Figure 18: Prediction of amino acid biomarkers for a set of amino acid metabolic disorders from Shlomi *et al.* Rows represent metabolic disorders and columns represent amino acids. The causative gene's name is indicated on the left. Blue and red entries represent biomarkers that are predicted by our method to be elevated or reduced, respectively. Table entries marked in '+' or '-' represent elevation or reduction in the metabolite's concentration in biofluids according to OMIM, respectively.

The figure reports specific metabolite changes, specifically amino acids, for a set of genetic diseases with impacts on metabolic enzymes. Correct predictions are when a blue cell (predicted increase) is paired with a '+' (observed increase) or a red cell (predicted decrease) is with a '-' (observed decrease). It also shows

some false positives (predictions with no observations) and false negatives (no prediction for an observed change). Overall, it serves well as a proof of concept of the prediction of biomarkers but lacks exploration of a variety of diseases and metabolites, since it is restricted to amino acids and related disorders.

Thiele *et al.* went on to predict more metabolic biomarkers for a better variety of IEMs [83], with the model Recon2 using FVA. The authors used a gold standard [131] of known IEM and associated biomarkers. The results were published in the form of the following heatmap in Figure 19b.

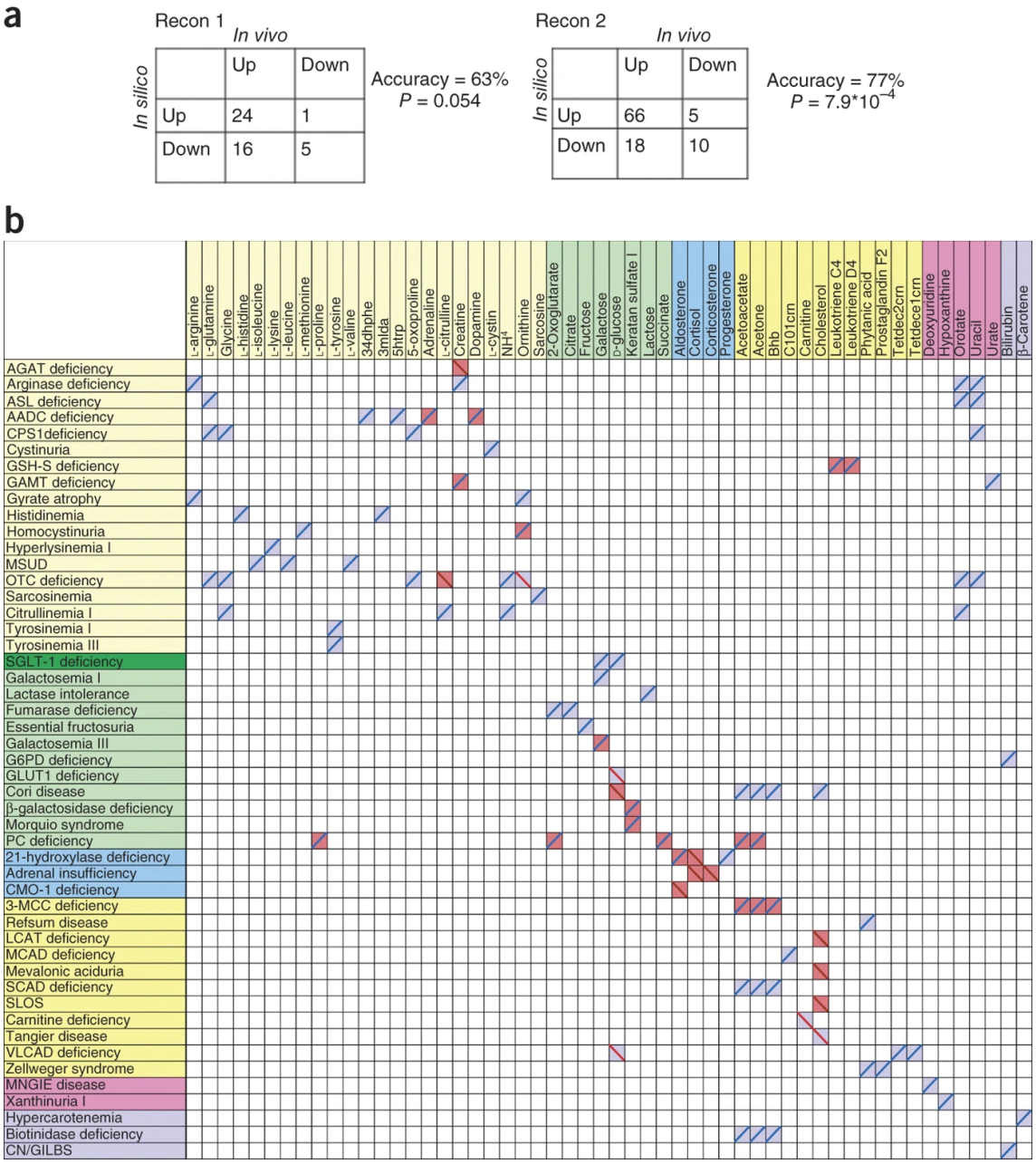


Figure 19: Predicted biomarkers for IEMs from Thiele *et al.*. (a) Comparison of the prediction accuracy of Recon1 and Recon2 against the gold standard [131]. (b) Correct and incorrect predictions. IEMs and biomarkers are sorted by subsystem. Bright yellow, amino-acid metabolism; green, central metabolism; blue, hormones; yellow, lipid metabolism; pink, nucleotide metabolism; lilac, vitamin and cofactor metabolism. Blue and red shading corresponds to predicted increase and decrease in biomarker, respectively. Blue and red lines represent reported increase and decrease of the biomarker in plasma, respectively.

This figure uses the same presentation as the previous heatmap in Figure 18 but with coloured slashes instead of '+' and '-'. By expanding both the number of simulated diseases and predicted biomarkers, the authors achieved a wider

variety of predictions across multiple metabolic classes. However, a major difference is the choice to remove all false positives and false negatives, meaning that predictions with no associated observation are not shown, and neither are observations with no predicted change. The choice to not show false positives could be explained by the fact that due to the nature of biomarker profiling, the other metabolites may have not been measured for all diseases and therefore lack any kind of information on whether there should be an observed change. However, the authors do not explain this choice in the paper and therefore the reasons for not displaying them are unknown. Additionally, they used this heatmap to calculate an accuracy score when comparing predictions with Recon1 (Figure 19a). In this case, the false positives were predictions of the opposite change direction compared with the observations. This means that the calculated accuracy score does not represent the ability to predict a biomarker or not, but the ability to correctly predict the direction of change.

1.2 Reproducing Thiele *et al.*'s predictions using FVA and random sampling

To investigate this lack of false positives and negatives from Figure 19, I reproduced the previous results using the Matlab code provided with the paper. The resulting heatmap, using the exact same metabolic model and parameters, is shown in Figure 20.

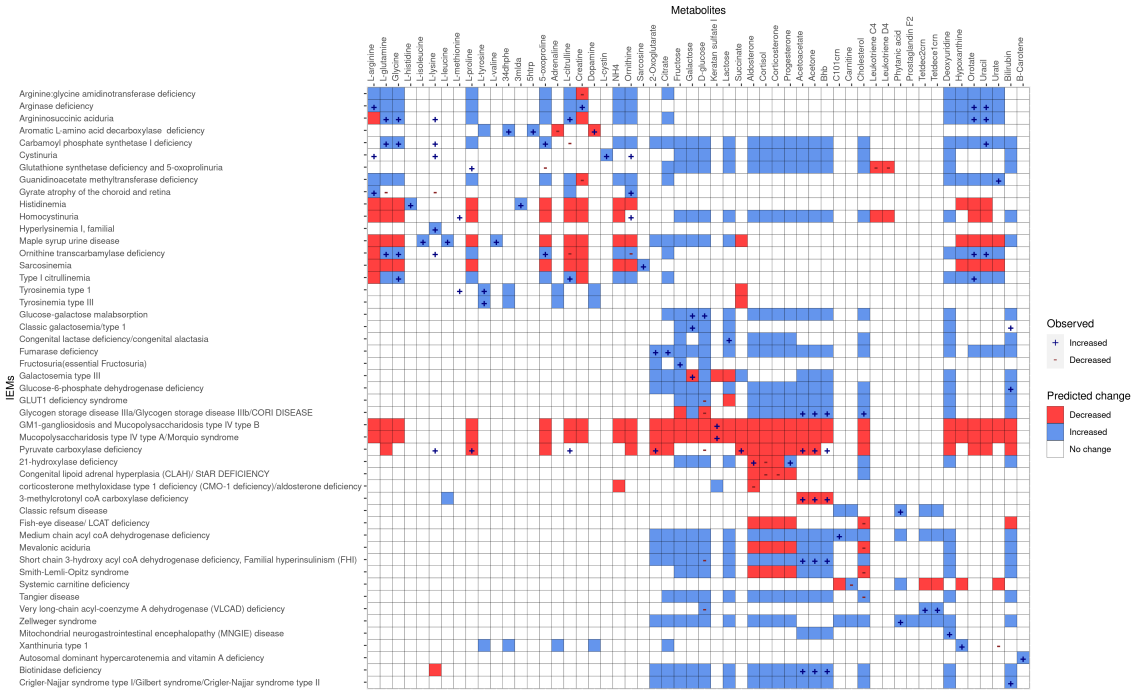


Figure 20: Reproduction of Thiele *et al.*'s heatmap using Matlab. Blue and red tiles represent biomarkers that are predicted to be increased or reduced, respectively. Table entries marked with '+' or '-' represent elevation or reduction in the metabolite's concentration in biofluids according to the gold standard [131], respectively.

In this figure we can clearly see the missing false positives and negatives that were not shown in Figure 20. The false positives make up a large portion of the predictions and remain important to validating the predictions and describing their accuracy.

Following this, I developed the same GSMN FVA-based biomarker prediction in Python to provide better reproducibility and availability for the community, and to make sure that the method was portable across programming languages. I coded the methodology from the Matlab script in Python using the cobrapy [130] package. I then reproduced the previous example, again using the same metabolic model Recon2, while attempting to match the parameters used in Matlab as close as possible. The resulting heatmap is shown in Figure 21.

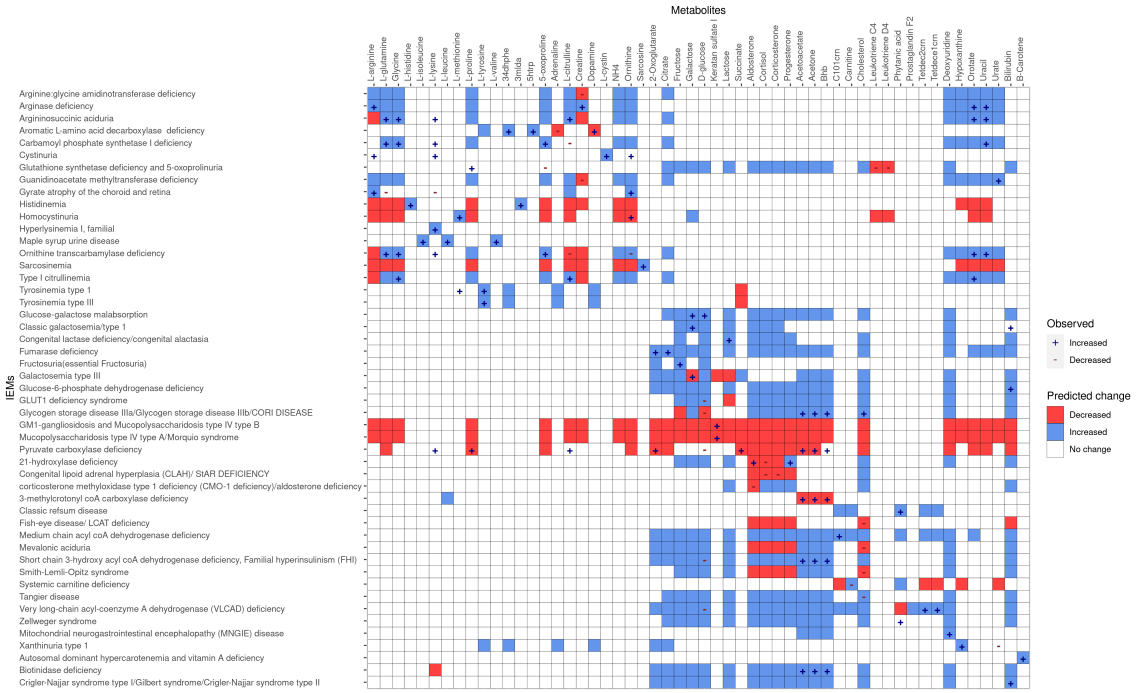


Figure 21: Reproduction of Thiele *et al.*'s heatmap using Python. Blue and red tiles represent biomarkers that are predicted to be increased or reduced, respectively. Table entries marked with '+' or '-' represent elevation or reduction in the metabolite's concentration in biofluids according to the gold standard [131], respectively.

There are some differences for certain diseases due to the inability to match certain parameters used in Matlab with Python, mainly to do with how the network is stored as Matlab (more matrix-oriented) versus Python (more dictionary-oriented) objects, and the corresponding value thresholds which are applied differently due to these format differences. The majority of predictions are identical to those from the Matlab simulations. This also serves as a "sanity check" since the false positives and negatives are still present, just as with the Matlab version.

Next, I took the prediction method in a different direction in an attempt to reduce the biases induced by the FVA approach to predicting metabolic biomarkers. In order to do this, I developed SAMBA, described in Section 3, which integrates random sampling with the previous biomarker prediction methodology. Once SAMBA had been sufficiently developed, I reproduced the same heatmap with random sampling and z-scores instead of the previously

used FVA methodology, as shown in Figure 22. The metabolite z-scores are calculated as described in Section 2.4.1 and quantify not only the intensity but also direction of predicted change between healthy and disease conditions.

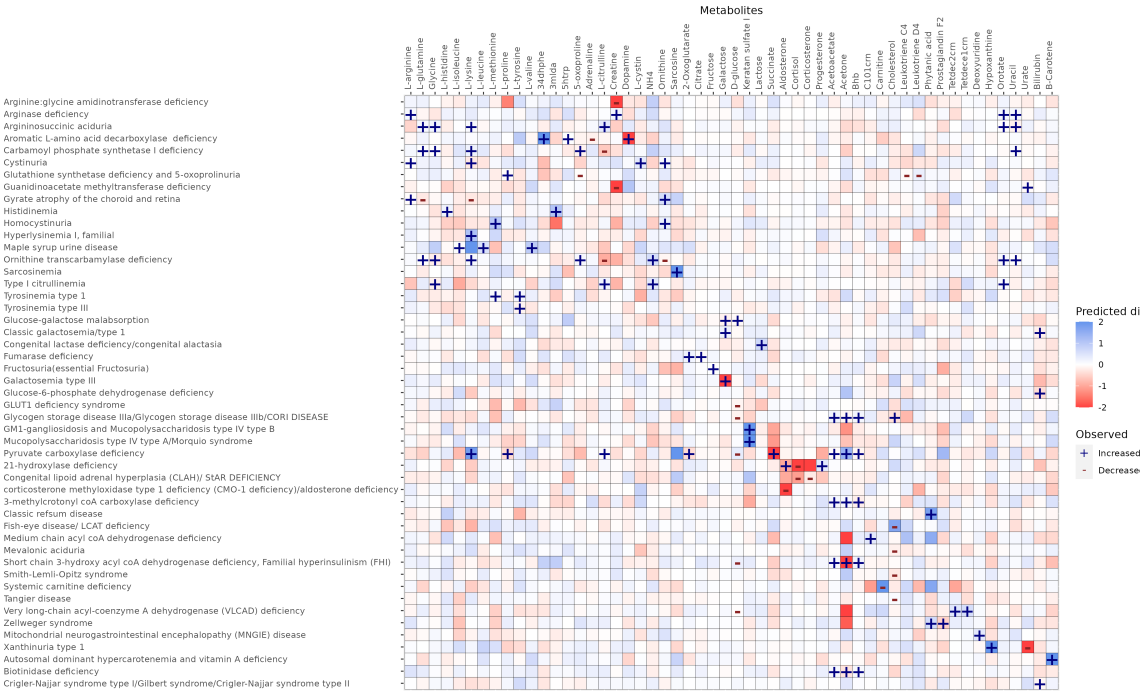


Figure 22: Reproduction of Thiele *et al.*'s heatmap with random sampling. The blue to red tile gradient represents biomarkers that are predicted to be more or less increased or reduced, respectively. Table entries marked with '+' or '-' represent elevation or reduction in the metabolite's concentration in biofluids according to the gold standard [131], respectively.

The sampling results, due to the use of no cut-off, show a gradient of z-scores which is more difficult to evaluate systemically than the previous binary results. Despite this, some observed biomarkers are correctly predicted, while others are not. This view of the metabolite predictions also highlights the scale of changes as opposed to the binary "on/off" changes shown in previous heatmaps.

To more easily compare between the two predictions, Figure 23 shows the overlap of both FVA- and sampling-based biomarker predictions using the same example dataset as the previous heatmaps, in the form of Venn diagrams.

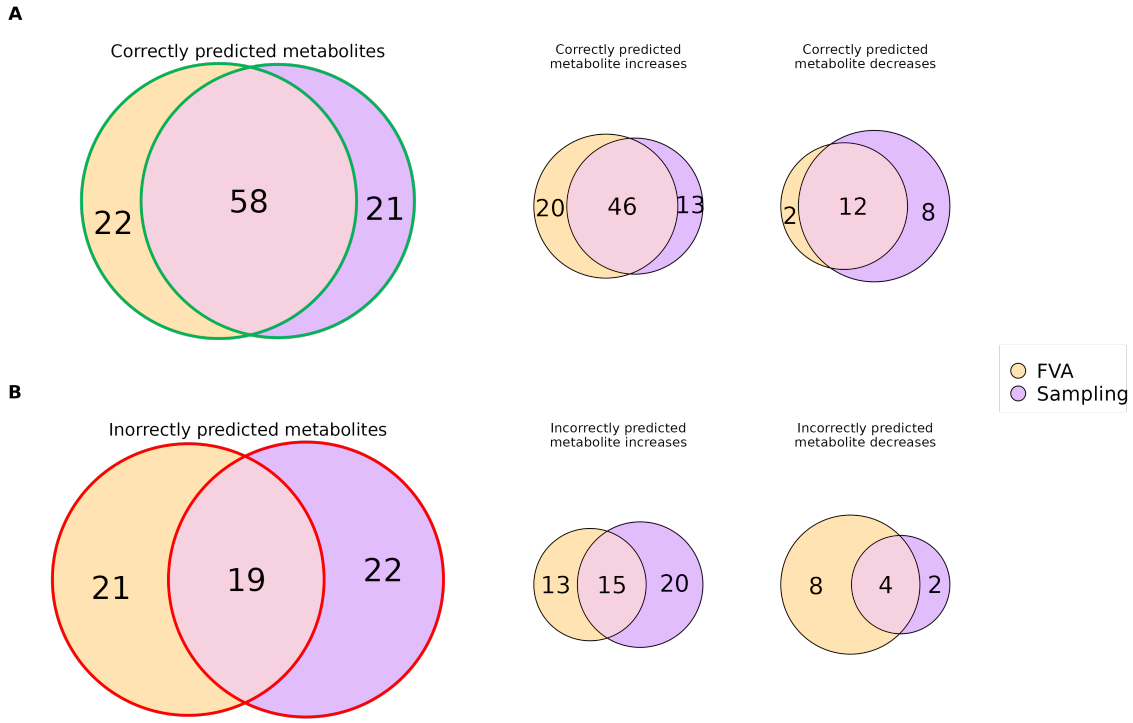


Figure 23: Venn diagrams of FVA (orange) and sampling (purple) predictions using Recon2. This shows the overlap between various predictions of the IEM subset using FVA and sampling. In orange, the FVA predictions are classed as correct or incorrect depending on the expected increase, decrease or no change. In purple, sampling was run on the same model with the same reaction KOs and model parameters. No z-score threshold was used, meaning that no metabolite is truly classed as "no change" for sampling, and only the prediction of directionality was taken into account. A correctly predicted increase (decrease) is a predicted increase (decrease) where the IEM subset reports an observed increase (decrease respectively). An incorrect increase or decrease is the sum of (i) predictions of the incorrect change direction (for FVA and sampling) and (ii) missing predictions where an IEM change is observed (for FVA).

Regardless of prediction change direction, FVA and sampling both predict a similar number of correct and incorrect metabolite changes, with each method predicting new metabolite changes not predicted by the other. This comparison highlights the difficulties in comparing the differences in each method, as FVA is used with a binary conversion to increase/decrease/no change, whereas with sampling a gradient of scores highlights the changes in many metabolites at once. More specifically, while the predictions of the two methods overlap frequently, sampling predicts metabolite decreases better than FVA: 20 correct predictions vs 14 for FVA, and 6 incorrect predictions vs 12 for FVA, but predicts fewer metabolite increases correctly.

A different method of displaying the sampling results is to plot the ranks directly on the heatmap instead of the z-scores as in Figure 24. The ranks are calculated by ordering metabolites by decreasing absolute value of z-scores, which results in the most deregulated metabolites, regardless of direction, being at the top of the ranked list. This of course causes the loss of information on the direction of each change but can serve to quickly highlight the most deregulated metabolites, especially when filtering using the top 10 most deregulated metabolites per condition for example (Figure 25).

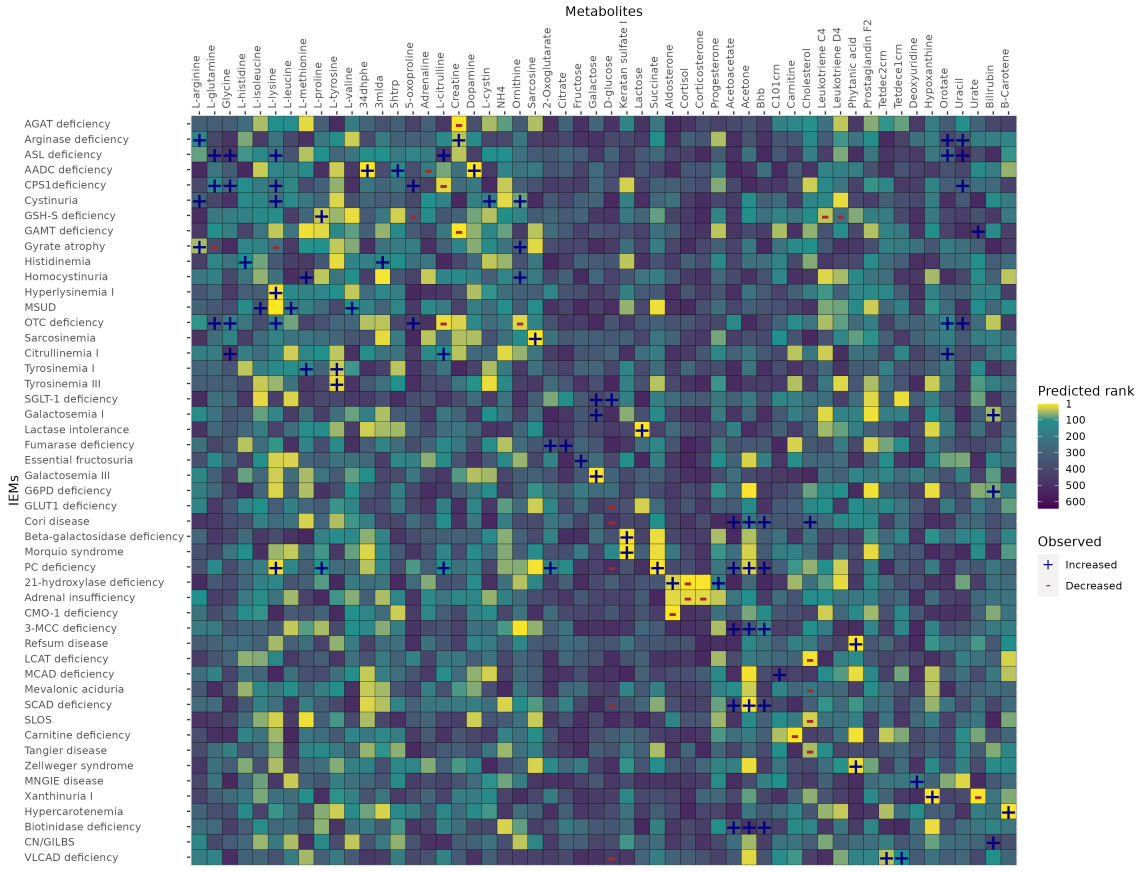


Figure 24: Heatmap based on Figure 22, using predicted ranks instead of z-scores, showing all ranks for each condition.

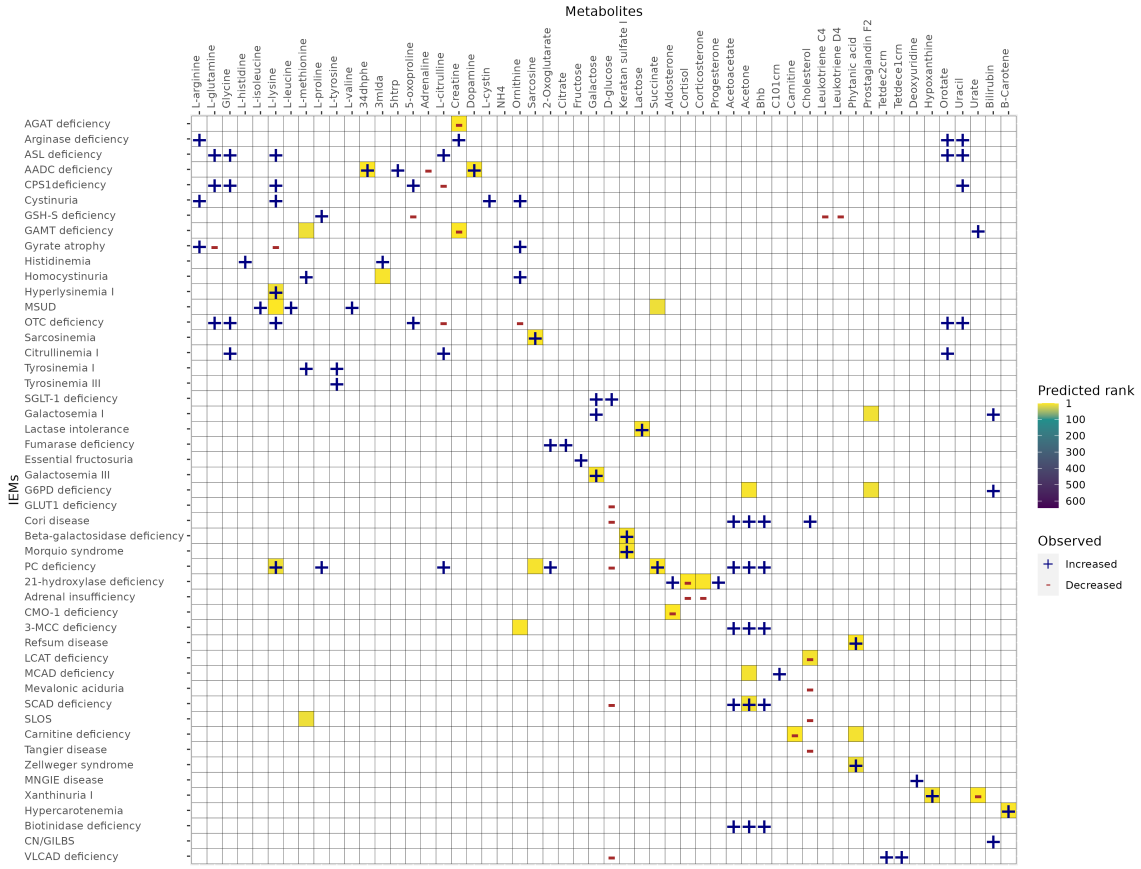


Figure 25: Heatmap based on Figure 22, using predicted ranks instead of z-scores, showing metabolite ranks with a cut-off of 10. Each row shows the metabolites in the top 10 for that condition. Most rows have less than 10 metabolites due to the fact that the figure only shows 54 of the ∼600 exchange metabolites in the model.

The use of sampling instead of FVA of course begs the question: why use random sampling for the prediction of metabolic profiles? The following sections demonstrate the advantages of sampling using different models and different experimental data, and serve as applications of the entire methodology including the use of z-scores and metabolite ranks.

2 Illustrating the benefits of sampling through the prediction of Xanthinuria type I biomarkers

Xanthinuria type I is a rare genetic disease caused by a mutation in the XDH gene [140], and is characterised by kidney stones (urolithiasis), urinary

tract infections, and rarely kidney failure [141]. In patients with this disease, a decrease in urate and an increase in hypoxanthine has been observed (from OMIM [1]).

Here, we applied both FVA and sampling in order to compare the information which can be drawn from both techniques. The flux simulations were run using Recon2 [83], a human genome-scale metabolic network containing 7 440 reactions, by knocking out the XDH gene, which knocks out 7 reactions (see Table 3). The exact version of Recon2 used in this example can be found at https://github.com/opencobra/COBRA.papers/tree/master/2013_Recon2. Recon2 was used for the IEM analyses as the idea was to compare results between FVA and sampling for the same set of conditions in the same model. This served as a proof of concept and we decided to publish the results using Recon2 to show the comparison with previous work by Thiele *et al.*. Sampling and FVA were run using the same parameters as in Shlomi *et al.* and Thiele *et al.* [128, 83]: minimum fraction of optimum of the objective function (biomass) set to 0, and all exchange reaction bounds set to $[-1, 1000]$.

Condition	Model	Gene KO	Reaction KO
Xanthinuria type I	Recon2	XDH	XANDp XAO2x XAOx r0424 r0425 r0546 r0547

Table 3: Genes and reactions knocked-out to simulate Xanthinuria Type I in Recon2.

Both the FVA bounds and the sampling distributions are displayed on the same plot for both of the expected biomarker metabolites (Figure 26). Expected biomarkers are defined as metabolites with observed significant changes in patients with the disease according to the original dataset. The flux values are reported on a log scale for clarity, and y-axis is the distribution density, hidden for visual clarity as the density values are not important for calculating z-scores and comparing distribution shifts.

For urate, the FVA bounds were the same in both conditions (Figure 26A),

which is interpreted as a metabolite not considered as a biomarker in Shlomi *et al.* [128], and thus the FVA prediction does not agree with the observed decrease. On the other hand, the sampling distributions correctly show a decreasing shift from WT to MUT. For hypoxanthine (Figure 26B), both methods are able to predict the expected increase in metabolite export via the shift in distributions for sampling and the change in upper bounds for FVA, although this change is small (10.3%) compared to the total feasible range.

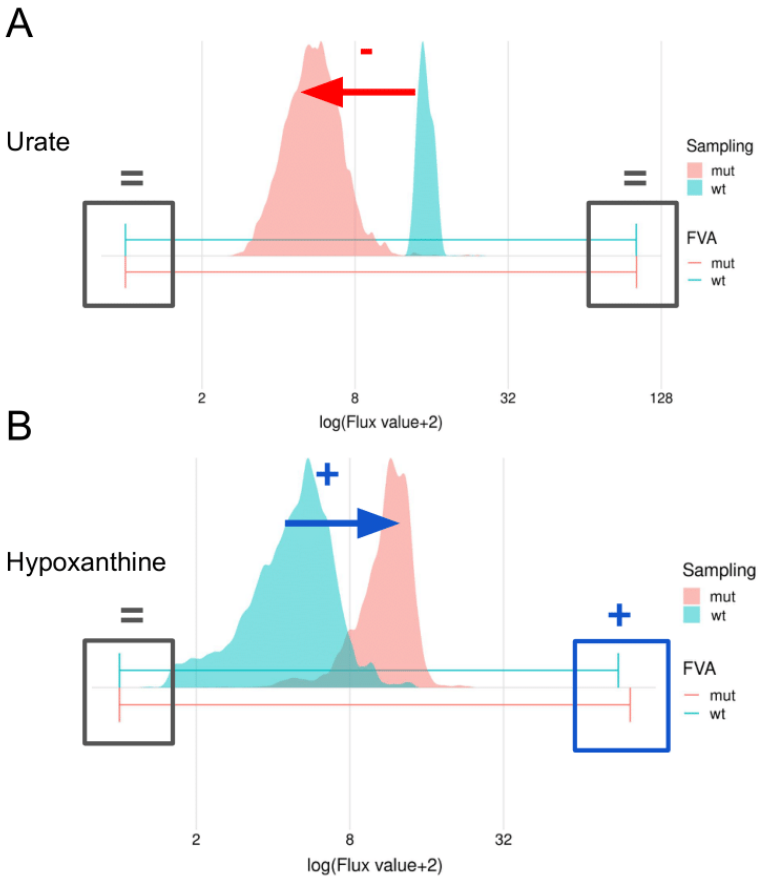


Figure 26: Flux bounds (FVA) and distributions (sampling) for urate (A) and hypoxanthine (B) in the WT state (light blue) and the MUT state (red). MUT here corresponds to the knock-out of the XDH gene. Highlighted in grey, red, and dark blue are the absences of shifts (=), decreases (-), and increases (+) respectively between WT and MUT.

Xanthinuria type I is one of many IEMs from an entire IEM - biomarker dataset which was curated in Sahoo *et al.* [131]. We used a subset of this dataset, used in Thiele *et al.* [83], to run our analyses by knocking out each gene responsible for

each disease: we ran both FVA and sampling on the 49 IEMs for 54 metabolites on Recon 2. Heatmap figures containing the entire set of predictions for both FVA and sampling are included in Figure 22 and Figure 24, and overlaps are shown in Figure 23.

Overall, sampling not only complements FVA by providing new correct predictions, but also attributes more meaning to the scores of the predictions for each metabolite (see Figure 22). Sampling z-scores, as opposed to the binary increase/decrease indicators of FVA, can be used to rank, filter and gain insight on the intensity of changes. Indeed, once generated, distributions can be used in multiple ways, one of which is calculating z-scores and ranks as shown in this paper; in contrast FVA interval boundaries represent only the flux extremes instead of the general flux behaviour.

3 Generating coherent recommendations and identifying novel metabolites of interest associated with SNPs using SAMBA

3.1 Introduction

For this study, we chose a case example for the comparison of mGWAS data and SAMBA ranking system. In general, GWAS datasets are composed of traits associated with SNPs, which are germline genetic substitutions of one nucleotide, present at a specific DNA position in at least 1% of the population. Specifically, mGWAS data consists of SNP-to-metabolic trait associations. One type of metabolic trait consists of single metabolite fold changes between non-SNP and SNP individuals (e.g. the fold change of margarate). The second type associated with SNPs in the study is ratios of two different metabolite levels, again compared between non-SNP and SNP individuals (e.g. the ratio

of margarate / palmitoleate). Other examples of these types of data are shown in Table 4 for illustrative purposes. We used data from Suhre *et al.* [142], extracted SNPs associated with significant metabolites, and mapped them onto the Human1 metabolic network. In this study, Human1 v1.10 [60], containing 13 024 reactions, was used to carry out all mGWAS analyses. It can be found at https://github.com/SysBioChalmers/Human-GEM. Human1 was used for the mGWAS analyses as we believe it is a more complete model, and it is in the community's best interest to use the latest model since it can then be improved by community efforts. It also highlights that SAMBA can scale to a larger model. Note that model choice will have an impact on any modelling approach and this selection step, out of the scope of this thesis, has to be taken with care. Sampling and FVA were run using the same parameters as in Shlomi *et al.* and Thiele *et al.* [128, 83]: minimum fraction of optimum of the objective function (biomass) set to 0, and all exchange reaction bounds set to $[-1, 1000]$.

Ratio	beta' meta	P meta	p-gain meta
myristate (14:0) / myristoleate (14:1n5)	0.124	$2.9 * 10^{-57}$	$1.2 * 10^{48}$
myristate (14:0) / palmitoleate (16:1n7)	0.131	$1.4 * 10^{-48}$	$1.0 * 10^{39}$
margarate (17:0) / palmitoleate (16:1n7)	0.157	$2.1 * 10^{-42}$	$6.6 * 10^{32}$
margarate (17:0)	0.06	$4.9 * 10^{-08}$	1
myristoleate (14:1n5)	-0.075	$3.3 * 10^{-09}$	1

Table 4: Examples of mGWAS data: 3 significant metabolite ratios and 2 significant single metabolites. Beta' represents the relative difference per copy of the minor allele (SNP) for the metabolic trait compared to the estimated mean of the non SNP population. The p-gain statistic quantifies the decrease in P value for the association with the ratio compared to the P values of the two separate corresponding metabolite concentrations.

Among the 37 SNPs present in Supplementary Table 3 of Suhre *et al.*, 17 were SNPs of 17 metabolic genes (one SNP per gene) present in the metabolic model Human1 (version 1.10). The 20 other SNPs were impossible to simulate since they do not correspond to metabolic genes in Human1. Human1 was used to run sampling on 2 of these 17 SNPs: SNPs affecting the Stearoyl-CoA 9-desaturase (SCD) gene and the Acyl-CoA Dehydrogenase Short chain (ACADS) gene. Human1 is one of the most recent and largest reconstructions of the human

metabolic network, also showing that the method can scale to this larger network (13 024 reactions, 8 363 metabolites). The 15 remaining SNPs with corresponding genes in Human1 were not analysed due to the manual curation needed to confirm genetic, enzymatic and metabolic matches.

We chose to focus on the SCD SNP specifically because i) the gene and reactions are present in the network, and ii) there are many measured metabolites present in the network, which is not the case for all of the SNPs, as some SNPs only have one or two significantly associated metabolites, or the associated metabolites do not exist in the network. It therefore serves as a good proof of concept application for the methodology. Furthermore, the selection of the correct genes to KO in the model for each SNP requires manual curation to make sure the GPR (gene-protein-reaction) relationships correctly represent the enzyme and corresponding gene. Additionally, mapping the metabolite names from the study to model metabolites is a time consuming manual step. Results for SCD are shown in the following figures in Section 3.2 and Section 3.3.1, and those for the ACADS SNP are shown in Section 3.3.2.

In contrast with IEM data, where mutations always result in an enzyme defect, an SNP might reduce enzyme activity (knock-down), enhance enzyme activity, or have an effect on a different gene. Some of the SNPs from the Suhre *et al.* study are well known to be associated with loss-of-function phenotypes such as enzyme deficiencies (e.g. the ACADS gene in ACADS-deficiency), and others have not been studied enough to confirm the effect of the SNP on gene function. As one example of an understudied SNP phenotype, the SCD gene (SNP rs603424 [143]) codes for the enzyme Stearoyl-CoA 9-desaturase, involved in fatty acid metabolism. The hypothesis is that the SNP mutation in the gene affects the corresponding enzyme negatively, which leads to no SCD enzyme activity, represented in the network by knocking-out the SCD gene and therefore blocking the corresponding reactions. This is suggested in Illig *et al.* [143] by drawing a parallel between known loss of function SNPs leading to severe

disorders, and newly identified SNPs. Additionally, the SNP mutation in the SCD enzyme-coding gene is predicted to be in an intronic (i.e. non-coding) region, using ensembl's VEP (Variant Effect Predictor) [144]. When simulating a scenario, the effect of the gene mutation should always be checked in order to generate the most accurate metabolic condition possible.

In Human1, there are 19 reactions linked to the SCD gene, most of which involve the desaturation of stearoyl-CoA, palmitoyl-CoA and myristoyl-CoA into corresponding mono-unsaturated fatty acids. Following the GPRs relationships in the model, knocking out SCD only affects 4 reactions (due to the fact that SCD can be compensated by another gene in the 15 other reactions, one of which is a transport reaction shared with 34 other genes). However, SCD also shares 14 GPRs with two other genes: SCD5 and FADS6, whose functions are not well described. We decided to knock out these extra 14 reactions in order to block the enzymatic function related to SCD completely (Table 5).

In the case of SCD, the GPRs were manually checked. The SCD SNP only affects the SCD1 gene (known as SCD in the metabolic model), as SCD1 and SCD5 are two separate genes. SCD5 codes for the same enzymatic function as SCD1 but they are both expressed in different tissues: fat tissue for SCD1, and brain and pancreas for SCD5. However, Human1 is not tissue-specific and the reactions are not necessarily associated with the genes according to this tissue specificity, so in order to block the enzymatic function completely, both SCD1 and SCD5 were blocked.

Condition	Model	Gene KO	Reaction KO
SCD	Human1 (v1.10)	SCD SCD5 FADS6	MAR02281 MAR02282 MAR02284 MAR02286 MAR02287 MAR02292 MAR02293 MAR02294 MAR02295 MAR02296 MAR02288 MAR02289 MAR00144 MAR00146 MAR00147 MAR00148 MAR02126 MAR02128

Table 5: Genes and reactions knocked-out to simulate SCD in Human1.

The SCD SNP has two types of significantly associated metabolic traits: single metabolite changes, and ratios of two different metabolite concentrations. The single significant metabolites measured for the mGWAS study for this SNP are margarate, palmitoleate, myristoleate, stearate and 1-palmitoleoylglycerophosphocholine. These are the main "expected" metabolites, which will be compared with the SAMBA recommended metabolites.

SAMBA returned z-scores for the 1497 unblocked metabolite exchange reactions in Human1. The distribution of these z-scores is shown in Figure 31, and highlights the difference between the extreme high-ranking metabolites and the low-ranking metabolites in the centre. A metabolic profile this large is difficult to compare with the data from the mGWAS study as no raw data was included in the original study: only the significantly associated metabolites were reported, as well as the total list of 295 measured metabolites (but not their fold changes for each SNP). We also calculated the FVA bounds for each metabolite for the same metabolic condition as the sampling. The results for the single metabolites and ratio metabolites are described in the following sections. Resulting SAMBA metabolites were manually mapped to the mGWAS significant metabolite names for SCD, with manual verification of metabolite synonyms as many lipids have multiple names and naming conventions.

3.2 Significant single metabolites for SCD

Here, we compared the 5 significant metabolites reported in the mGWAS study with their simulated SAMBA metabolite ranks and FVA bounds to see the biggest effect this KO has on metabolite exports and imports. No rank or z-score thresholds were used for Figure 27, Figure 28, and Figure 29 as the metabolites were selected based on their presence in the significant results of the Suhre *et al.* dataset.

Figure 27 shows the five metabolites identified in the mGWAS study along

with the corresponding SAMBA ranks and the FVA predictions. In Figure 27, Figure 28 and Figure 29, the metabolite(s) marked with "NA" in the SAMBARank column have no flux values because either they are not present as a metabolite in the network, don't have an exchange reaction in the network, or have a blocked exchange reaction, meaning no flux can be carried through it in the current metabolic state.

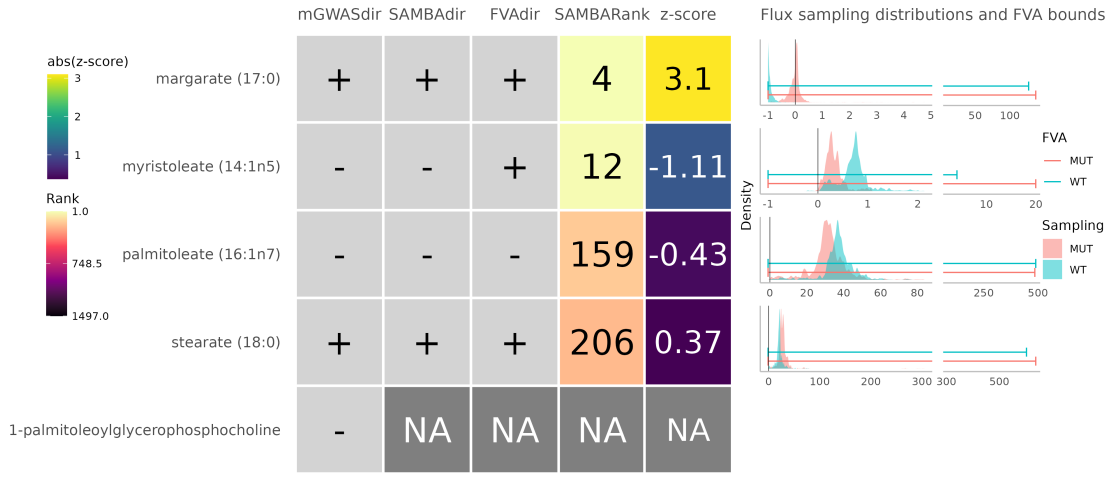


Figure 27: Observed and predicted changes for the five metabolites significantly associated with the rs603424 SNP. The first column shows the observed change directions from the mGWAS study. The second column shows the predicted change direction using SAMBA (SAMBAdir). The third column shows the predicted change direction using FVA (FVAdir). The fourth column shows the SAMBA predicted rank out of the 1497 metabolites in the network (SAMBARank). The fifth column shows the SAMBA predicted z-score, with the colour scale as the absolute value of the z-score. The NAs represent metabolites for which SAMBA was unable to predict fluxes for one of the following reasons: (i) the metabolite is not in the network, (ii) the metabolite is in the network but has no exchange reaction, or (iii) the metabolite's exchange reaction can carry no flux (=blocked). Sampling distributions and FVA predicted bounds for each metabolite's exchange reaction in WT and MUT are shown on the right.

Four out of the five expected metabolites are present with an exchange reaction in Human1, and the SAMBA predicted change directions match the expected mGWAS experimental changes. The directions of change predicted by FVA are correct except for myristoleate, which was predicted to be increased instead of decreased using the FVA bounds. Their ranks are shown in the column SAMBARank and these ranks are to be compared with the total number of exchange metabolites present in Human1, i.e. 1497. These four metabolites are

in the top 13%, two of which are in the top 1%. The z-scores are also shown in the last column, which highlights the difference in value between the top best ranked metabolites and the lower ranks.

3.3 Significant ratio metabolites

3.3.1 SCD ratios

The significant metabolite ratios linked to SCD include many different combinations of pairs of metabolites. The assumption here is that at least one of the two metabolites involved in each ratio must change for the ratio to be significantly changed. This is less direct than the previously shown significant metabolites, as they may not necessarily change as drastically between the two conditions, but they serve to extend the list of possible metabolites to map to using the SAMBA predictions. Figure 28 shows the metabolites present in at least one ratio significantly associated with SCD and their associated predicted SAMBARanks.

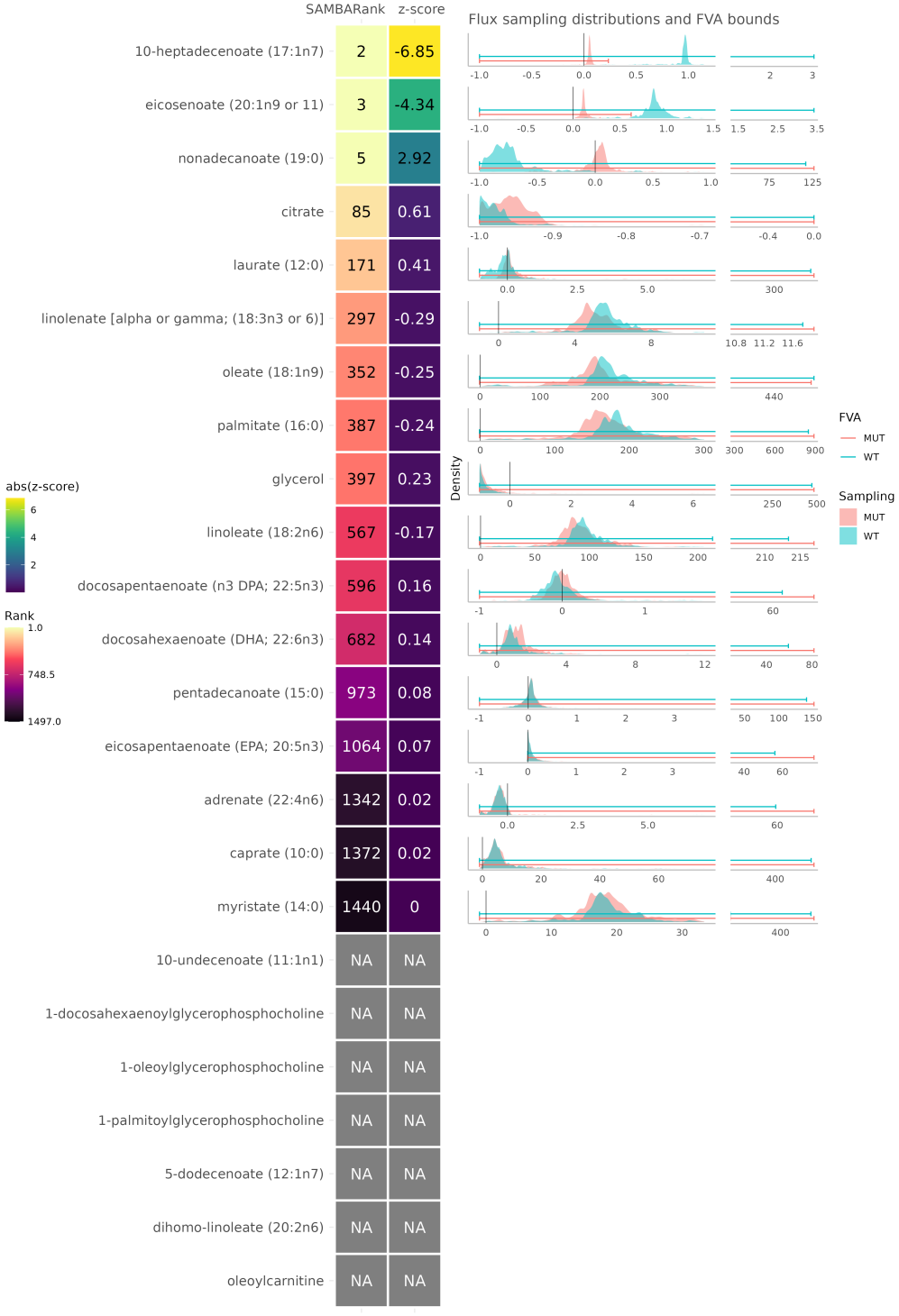


Figure 28: Predicted ranks for the metabolites present in a ratio significantly associated with the rs603424 SNP. The first column shows the predicted rank out of the 1497 metabolites in the network. The second column shows the SAMBA predicted z-score, with the colour scale as the absolute value of the z-score. The NAs represent metabolites for which SAMBA was unable to predict fluxes for one of the following reasons: (i) the metabolite is not in the network, (ii) the metabolite is in the network but has no exchange reaction, or (iii) the metabolite's exchange reaction can carry no flux (=blocked). Sampling distributions for each metabolite's exchange reaction in WT and MUT are shown on the right.

The second most differentially abundant metabolite predicted by SAMBA for this condition is 10-heptadecenoate, which is present in at least one significant ratio in the mGWAS SCD dataset. In addition to this, there are 4 other highly ranked metabolites, all in the top 171 ranked metabolites out of 1497 (top 11%). The five metabolites ranked below the 50% mark have z-scores lower than 0.1. Interestingly, myristate is almost ranked last in the entire list of predictions. When taking a closer look at its flux distributions, the MUT distribution appears to be bimodal, meaning that while the flux seems to have shifted, the z-score was not able to detect this difference due to its reliance on the similar means.

3.3.2 ACADS ratios

A second example from the Suhre *et al.* paper is the Acyl-CoA Dehydrogenase Short chain (ACADS) SNP. In the paper, it does not have any "single" metabolite trait associations, but has 11 significant ratio metabolite associations. The predictions using SAMBA for these metabolites involved in significant ratios are shown in Figure 29.

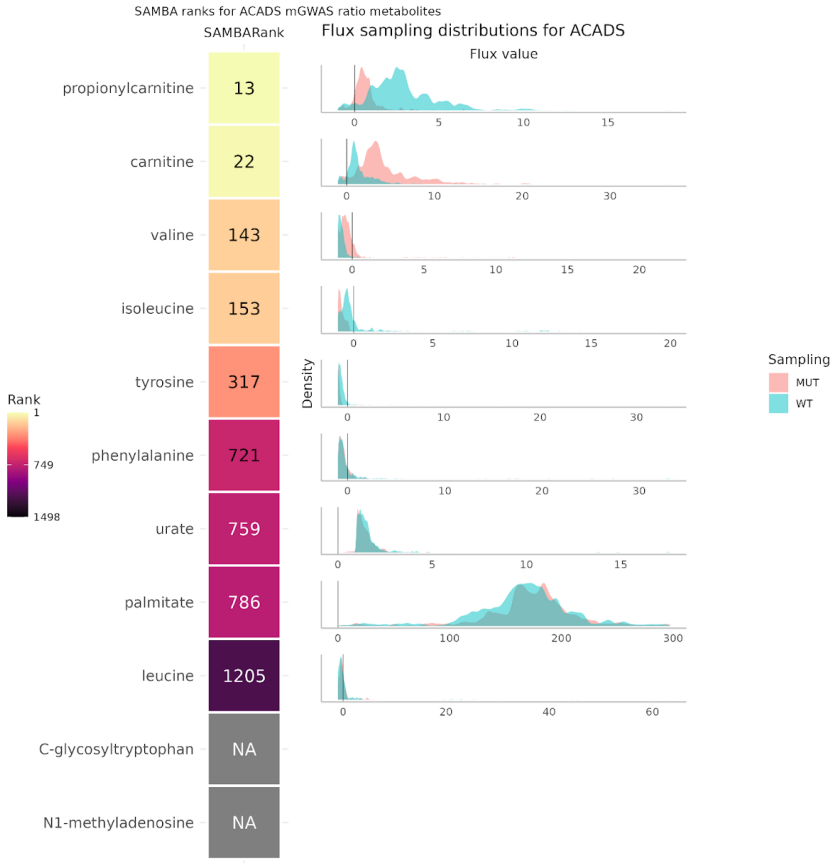


Figure 29: SAMBA ranks for the metabolites involved in significant ratios for the ACADS SNP from Suhre *et al.* 2011. The column shows the predicted rank out of the 1498 metabolites in the network. The NAs represent metabolites for which SAMBA was unable to predict fluxes for one of the following reasons: (i) the metabolite is not in the network, (ii) the metabolite is in the network but has no exchange reaction, or (iii) the metabolite's exchange reaction can carry no flux (=blocked). Sampling distributions for each metabolite's exchange reaction in WT and MUT are shown on the right.

As in the previous example, there are some extremely well ranked metabolites while others are poorly ranked. The literature suggests that carnitine (rank 22) is intrinsically linked with CoA on multiple levels. Metabolically, the reactions catalysed by ACADS enzymes are two reactions away from propionycarnitine and L-carnitine, due to their interaction with propanoyl-CoA, their direct product. Carnitine also plays a role in the stabilisation of CoA and acetyl-CoA levels, as well as energy production by taking part in a rate controlling step in mitochondrial oxidation of long-chain fatty acids [145]. Medically, L-carnitine is used as treatment in some cases of ACADS deficiency (also known as SCAD

deficiency (short chain acyl-CoA dehydrogenase)) [146]. Regarding the highly ranked amino acids, an adjacent enzyme Isobutyryl CoA Dehydrogenase (IBD), which is coded by ACAD8 and shares GPRs with ACADS, has been shown to be involved in valine metabolism [147, 148]. The ACADS gene is also involved GPRs in reactions in the "Valine, leucine, and isoleucine metabolism" pathway in the Human1 GSMN.

3.4 Significance of predictions

3.4.1 Statistical significance

Despite the problems that come with evaluating the false positives and negatives provided by the method, the statistical significance of the previous findings can be evaluated using a hypergeometric test. The test describes the statistical significance of predicting k number of metabolites correctly out of the top n predictions, when taking into account the total N number of predictions containing K number of experimentally significant metabolites.

Figure 30 shows the results of these tests for various rank cut-offs. For example, when looking at the top 300 (n) metabolites (x-axis), predicting 10 (k) experimentally significant metabolites (green y-axis) out of the 20 (K) total experimental metabolites for a total of 1497 (N) predictions, is significant (p-value < 0.05) (blue y-axis).

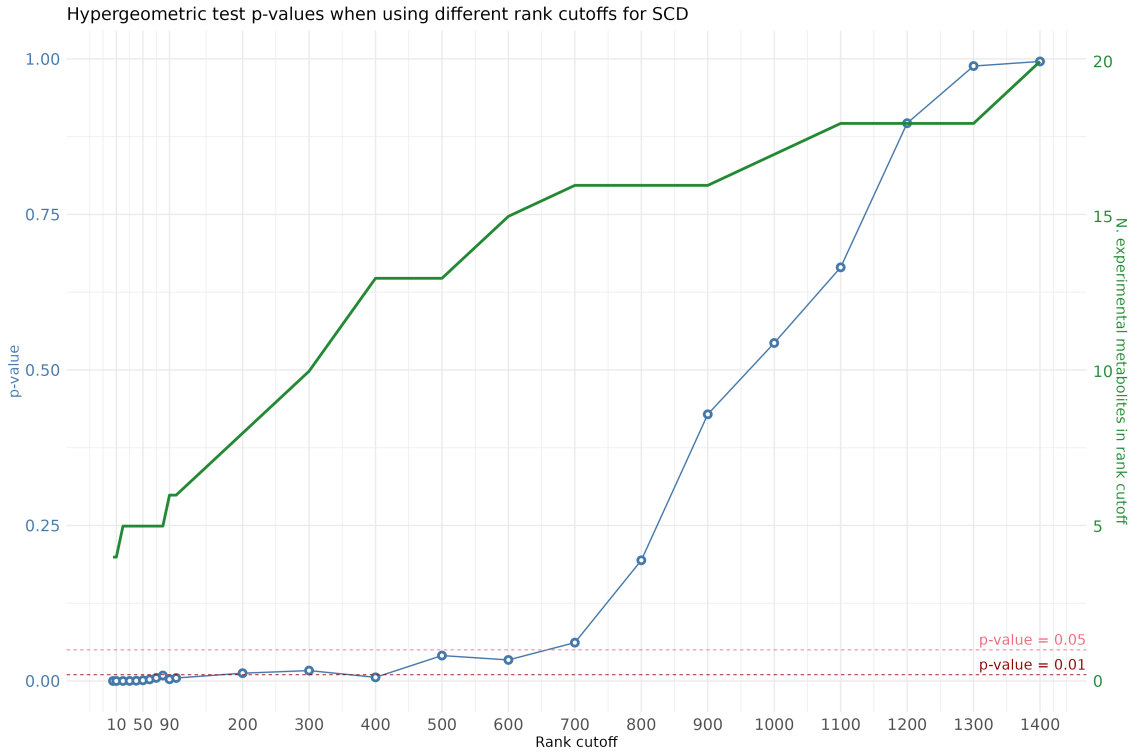


Figure 30: Hypergeometric test p-values for different rank cut-off values for SCD. The left y axis (blue) shows the hypergeometric test p-values when using a given rank cut-off and the number of experimental metabolites predicted in that top ranking. The right y axis (green) shows the number of experimental metabolites predicted for each rank cut-off.

The figure highlights the significance of finding theses numbers of expected metabolites in the top ranks of the SCD predictions. Until around the top 100, the test shows that predicting around 6 expected metabolites is extremely significant (p-value $\ll 0.01$) and remains significant (p-value < 0.05) until just below the halfway point of the ranked list.

3.4.2 Ranking provides the extreme metabolite changes

Metabolite ranks are determined from the highest absolute values of z-scores for each metabolite among the entire list of metabolite changes for that condition. This means that the top best ranked metabolites are relative to the rest of the list. Generally, the top most changed metabolites have ”extreme” z-score values due to large shifts in distributions caused by the metabolic perturbation. In the SCD example, Figure 31 shows the distribution of z-scores for all metabolite

predictions. Those with a z-score higher than 1 or lower than -1 are highlighted in blue (threshold chosen for illustrative purposes), and the expected metabolite names predicted in the top 10 are shown in red.

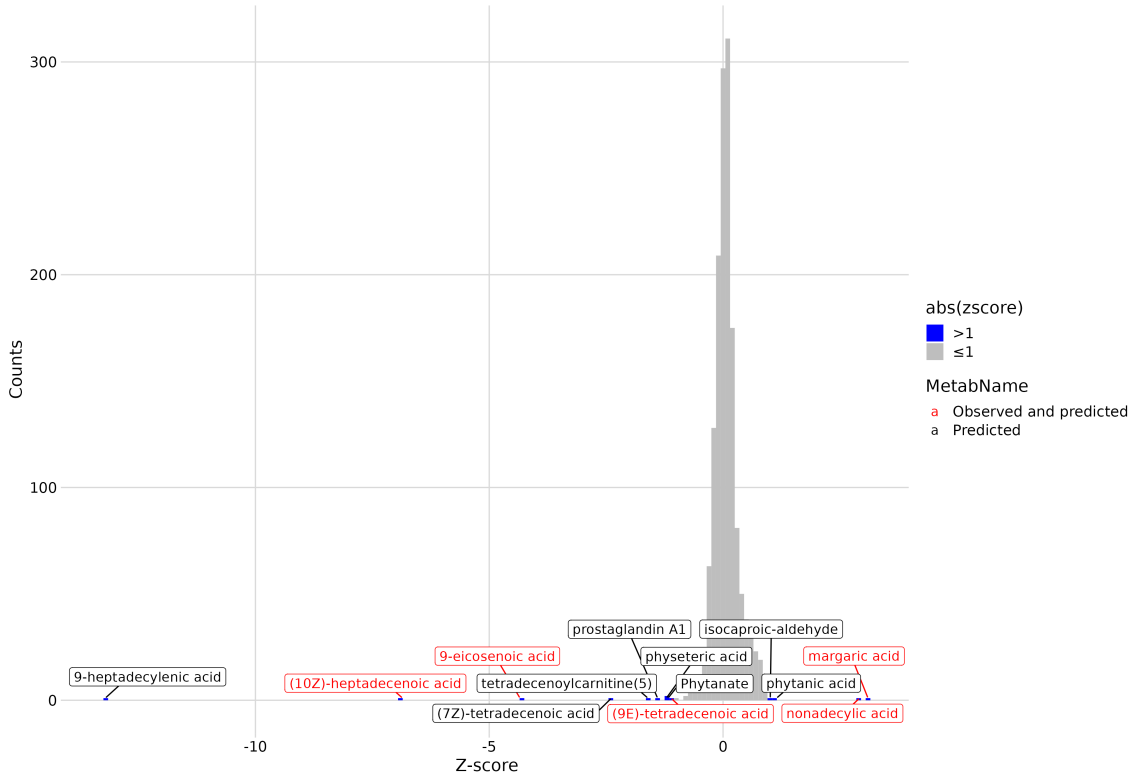


Figure 31: Distribution of metabolite z-scores for SCD. The metabolite labels highlighted in red are significantly observed in the Suhre *et al.* paper. Here, a threshold of 1 is used to show the metabolites that pass the threshold (blue).

This distribution shows how the few extreme z-scores are different to the main "body" of the distribution, and highlights the fact that differences between low ranks are very small in terms of z-score values. Most z-scores are very close to 0 meaning that the rank at this level does not hold much value when comparing between two low ranks. This of course must be taken into consideration when interpreting these simulated metabolic profiles.

3.5 Using SAMBA predicted metabolite lists can enrich experimental knowledge

The top most differentially changed metabolites associated with SCD predicted using SAMBA can be used to form a list of new metabolites of interest for this condition. By examining the chemical class of each predicted highly differentially abundant metabolite, we can gather information on a general type of metabolite affected by the KO. This section displays three approaches to bridging the gap between enriching predicted data and linking it to experimental data.

3.5.1 BiNChE, a ChEBI-based enrichment analysis for metabolites

BiNChE [149] creates and enriches a subnetwork using a list of ChEBI IDs and the ChEBI ontology. For this, I used the top 10 most changed metabolites as predicted by SAMBA for the SCD example presented previously. Table 6 shows this list of top 10 metabolites, and the rows shown in blue are the experimentally significant metabolites for the SCD SNP.

Metabolite name (alternate name)	Human1 ID	CHEBI	SAMBA rank
9-Heptadecenoic acid	MAM01238e	80550	1
10-Heptadecenoate	MAM00003e	75094	2
gadoleic acid (eicosenoate)	MAM01235e	32419	3
heptadecanoic acid (margarate)	MAM02456e	32365	4
nonadecanoate	MAM02613e	39246	5
cis-tetradec-7-enoic acid	MAM00117e	53206	6
(5E)-tetradecenoyl-L-carnitine	MAM02974e	131957	7
prostaglandin A1	MAM02776e	15545	8
5-Tetradecenoic acid (physeteric acid)	MAM02745e	89393	9
Phytanic acid	MAM02746e	16285	10

Table 6: Table of the top 10 most differentially changed metabolites for the SCD gene KO using SAMBA. Human1 IDs correspond to the exchange metabolite IDs. Metabolites highlighted in blue were also significant in the SCD mGWAS data.

Using this list as input for enrichment, Figure 32 shows a subnetwork of the ChEBI ontology. BiNChE is a good alternative to classic pathway enrichment

analyses (gene set enrichment analysis (GSEA)/metabolite set enrichment analysis (MSEA), overrepresentation analysis (ORA)) due to its use of the richer structure of the ChEBI ontology as opposed to the approach of focusing on metabolites and pathways of interest.

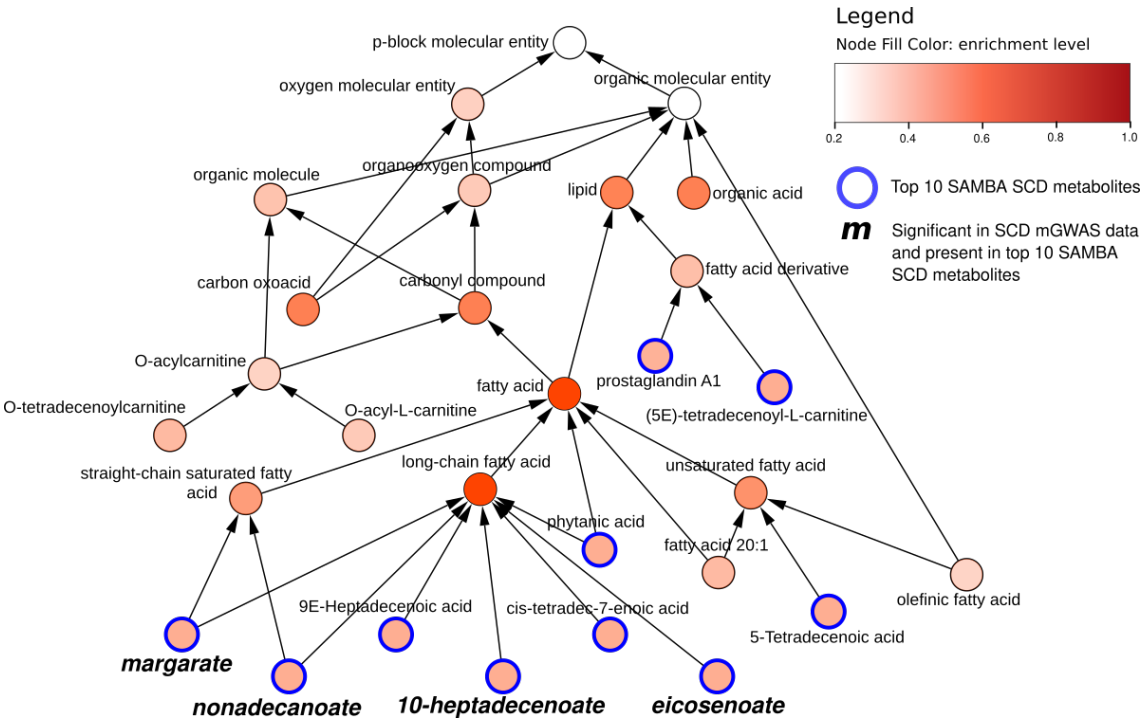


Figure 32: Hierarchical ChEBI graph of the top 10 metabolites predicted to be differentially abundant (outlined in blue) by SAMBA for SCD, extracted using BiNChE. The node colour corresponds to the BiNChE enrichment level. The metabolites in bold & italic were significant in the mGWAS dataset for SCD.

All of the top 10 most changed metabolites are classified as lipids (outlined in blue in Figure 32), 8 of which are fatty acids, which is consistent with the functionality of the enzyme SCD. Indeed, the SNP rs603424 has been shown to be significantly associated with circulating phospholipid levels [150], as well as with low levels of palmitoleate [150]. SCD is a desaturase which leads to the formation of fatty acids, specifically monounsaturated fatty acids involved in membrane phospholipids [151]. While this information could appear evident, it helps confirm SAMBA predictions, and it can reveal insight into metabolic classes which are not yet known by the user, serving as a good starting point for further

research into the chemical classes of interest for a given metabolic state.

Out of the top 10 metabolites, 4 were measured in the mGWAS study (margarate, 10-heptadecenoate, nonadecanoate, and eicosenoate), and they are all classified as saturated or long-chain fatty acids. This means that the other long-chain fatty acids could be potential metabolites of interest, such as 9-Heptadecenoic acid (rank 1) or cis-tetradec-7-enoic acid (rank 6), which weren't measured in the original mGWAS study.

However, the ChEBI classification is limited by the annotation of each metabolite to the correct class. Upon manual inspection, both cis-tetradec-7-enoic acid and 5-tetradecenoic acid are C14:1 fatty acids, only differing by the position of the double bond, but they are classified separately in long-chain fatty acid and unsaturated fatty acid respectively. This indicates that 5-tetradecenoic acid could also be of interest for future studies. Furthermore, by looking at the chemical structures, 4 out of the top 10 are odd chain fatty acids which is interesting to highlight since they represent a very small percentage of the total human fatty acid plasma concentration [152].

Since BiNChE provides a view of the ChEBI ontology on a per-metabolite scale, using too many metabolites as input results in a large and difficult to read figure. Other methods can integrate more of the predicted metabolic profile (for example 50 metabolites), such as the following approach.

3.5.2 ChemRich enriches chemical classes based on molecular data

As a step closer to using chemical structures as opposed to class annotations as well as using more of the metabolic profile, we ran a ChemRich [153] analysis using the top 50 metabolites predicted to be differentially abundant for the same SCD example as before. ChemRich uses the chemical structure via SMILES, and the MeSH terms associated with PubChem IDs to highlight enriched chemical classes. Figure 33 represents the most enriched clusters from the top

50 metabolite set. The higher the -log(pvalue) (y axis), the more the group is enriched. The x-axis serves to separate the groups for plotting purposes using a chemical similarity tree behind the scenes.

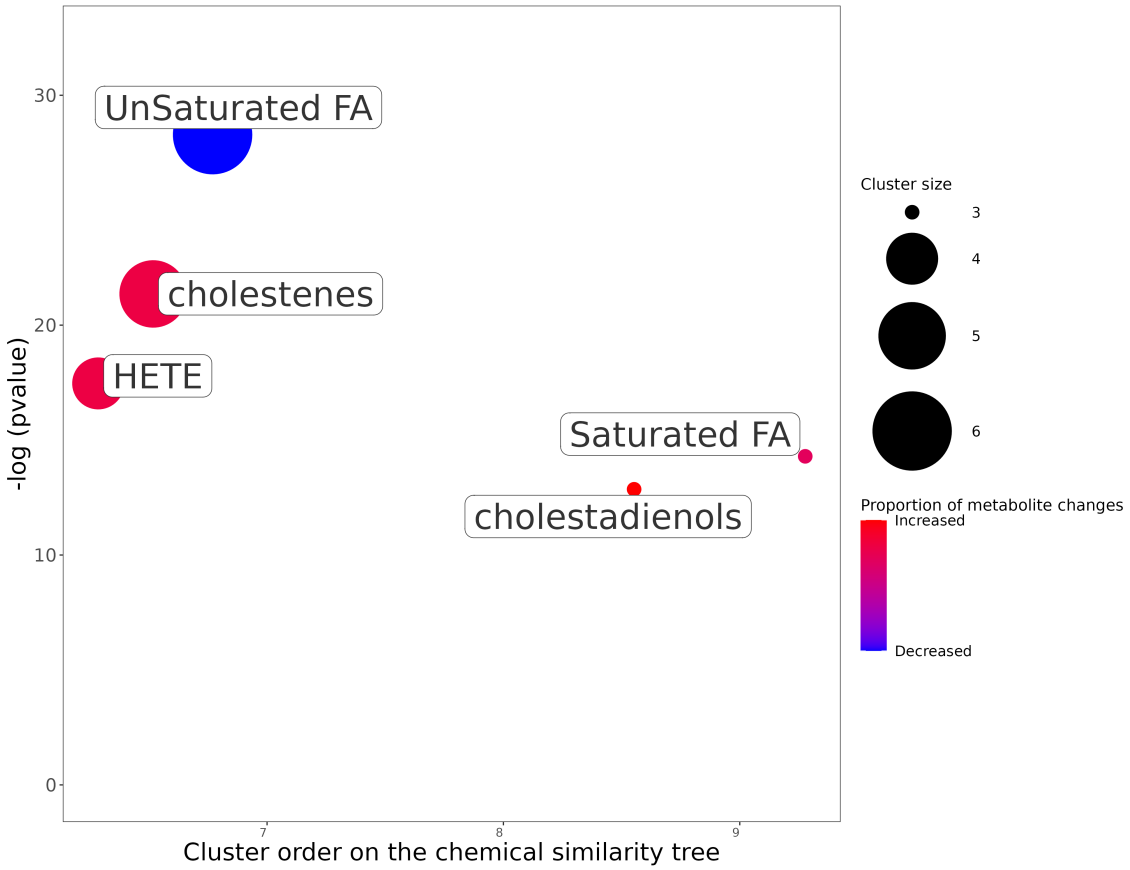


Figure 33: ChemRich enrichment of the top 50 most changed metabolites for SCD. The y-axis shows the most significantly altered clusters on the top. Each node reflects a significantly altered cluster of metabolites. Enrichment p-values are given by the Kolmogorov–Smirnov test. Node sizes represent the total number of metabolites in each cluster set. Cluster colours show the proportion of increased or decreased metabolites (red and blue respectively). The x axis represents a separation based on cluster order on the chemical similarity tree, and non-significant clusters are hidden.

The ChemRich plot also shows that both saturated and unsaturated fatty acids are significantly enriched by this dataset when including 40 more metabolites. Figure 33 also highlights some other groups such as HETE (Hydroxyeicosatetraenoic acids (which are oxylipins)), cholestenes, and cholestadienols not detected using BiNChE (which is most likely due to the fact that BiNChE was given 10 metabolites instead of 50). ChemRich serves

as a complsignificantlyementary method to BiNChE for analysing predicted metabolic profiles, as highlighted in Table 7.

Compound Name	SMILES	Z-score	Cluster
9-heptadecylenic acid	CCCCCCC/C=C/CCCC CCCC(=O)O	−13.211	UnSaturated FA
(10Z)-heptadecenoic acid	CCCCCC\C=C/CCCCC CCCC(O)=O	−6.854	UnSaturated FA
9-eicosenoic acid	[H]\C(CCCCCCCCCC)= C(/[H])CCCCCCCC(O)=O	−4.342	UnSaturated FA
margaric acid	CCCCCCCCCCCCCCCC (=O)O	3.099	Saturated FA
nonadecylic acid	CCCCCCCCCCCCCCCCCC (=O)O	2.920	Saturated FA
(7Z)-tetradecenoic acid	CCCCCC/C=C\CCC CCC(=O)O	−2.387	UnSaturated FA
tetradecenoyl -carnitine(5)	CCCCCCCC/C=C/CCCC (=O)O[C@H](CC(=O)[O-]) C[N+](C)(C)C	−1.606	
prostaglandin A1	CCCCC[C@@H](/C=C/ [C@H]1C=CC(=O)[C@@H] 1CCCCCCC(=O)O)O	−1.400	prostaglandins a
physeteric acid	CCCCCCCC/C=C/CCCC (=O)O	−1.213	UnSaturated FA
Phytanate	CC(C)CCCC(C)CCCC(C) CCCC(C)CC(=O)O	−1.175	Saturated FA

Table 7: ChemRich provides a metabolite-level table with each metabolite assigned to a cluster, the top 10 of which are shown in Table 7. These clusters are more specific than some of the annotations provided by ChEBI, namely those assigned to "long-chain fatty acid" by BiNChE are grouped into unsaturated and saturated fatty acids by ChemRich.

Additionally, when using only the top 20 SAMBA predictions as input for ChemRich, the same classes as those obtained when only using the list of 21 experimentally significant metabolites are identified, shown in Figure 34. By going further down the list of ranked predictions, the information gained can be enriched using the simulated data. This figure clearly shows the gain of information as the list of original metabolites grows in length and is enriched.

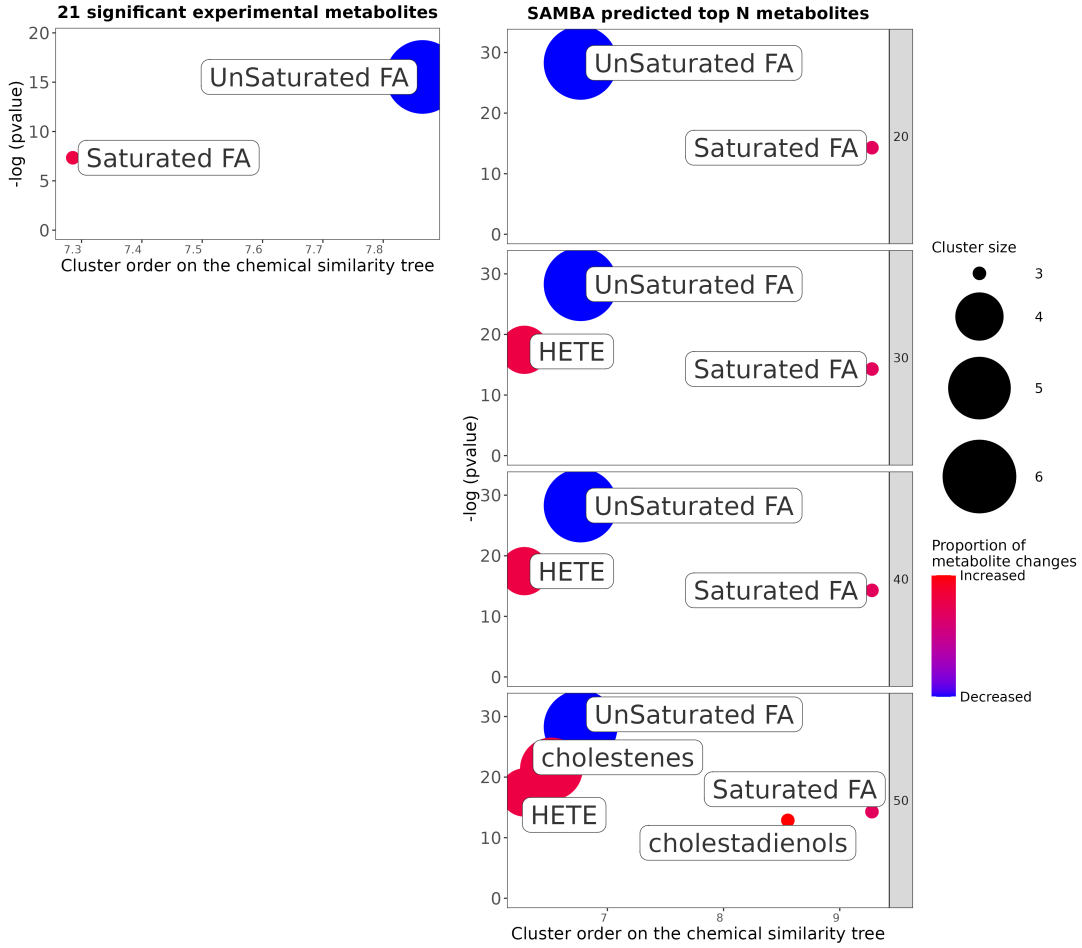
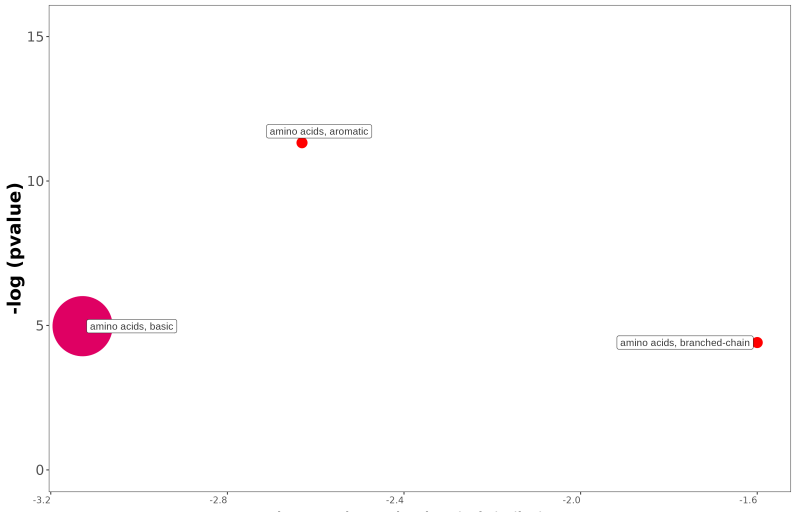


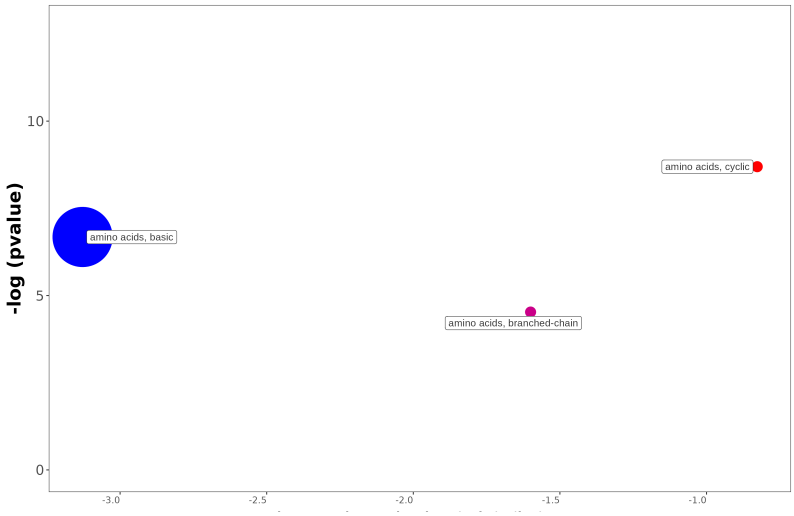
Figure 34: ChemRich using only experimentally significant metabolites (left) and using increasing numbers of highly ranked SAMBA predicted metabolites (right) for SCD.

Finally, I ran ChemRich on the IEM examples from the Thiele *et al.* study using the sampling predictions for each IEM (Figure 22), using only the metabolites present in the figure (all 54 columns). Figure 35 shows three ChemRich figures for three IEMs when using all 54 sampling metabolite predictions (z-scores) for each condition as input. The three IEMs are: Aromatic L-amino acid decarboxylase deficiency, Fish-eye disease/ LCAT deficiency, and Autosomal dominant hypercarotenemia and vitamin A deficiency. They are representative of the other 46 ChemRich figures (not shown here) in that they all show mainly amino acid enrichments even for non amino acid related diseases, due to the

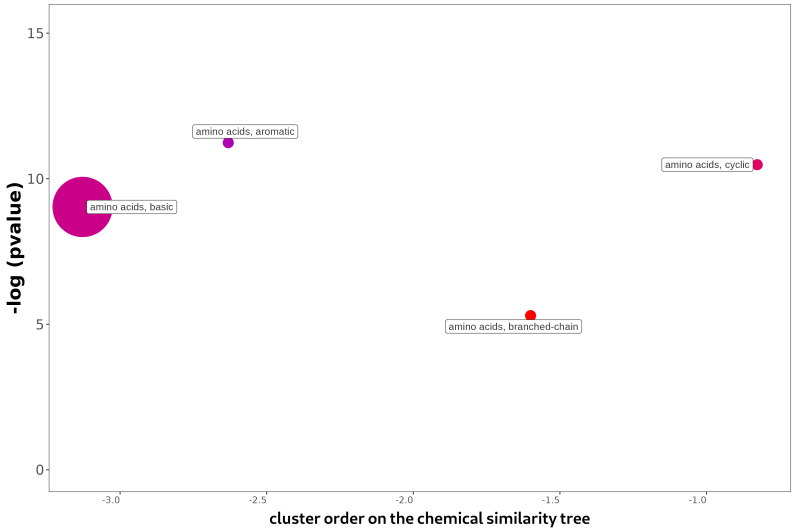
many amino acids present in the input metabolites (see Figure 19 for the classes of each metabolite and IEM shown as colours).



(a) Aromatic L-amino acid decarboxylase deficiency



(b) Fish-eye disease/ LCAT deficiency



(c) Autosomal dominant hypercarotenemia and vitamin A deficiency

Figure 35: ChemRich using sampling predictions for the 54 metabolites from Thiele *et al.* for three IEM conditions.

The resulting ChemRich figures in Figure 35 show that the enrichment is mainly influenced by the presence/absence of a metabolite in the list rather than its fold change or score, which led to the ChemRich enrichment always producing very similar enrichments (mainly enriched amino acids) when using the same list of 54 metabolites with different scores. This shows that by using a rank threshold, the initial bias from choosing which metabolites to predict for is disregarded since the entire list of metabolites is taken into account to create the list of top 50 most changed metabolites. For example, autosomal dominant hypercarotenemia and vitamin A deficiency is unrelated to amino acids, being more involved with vitamins and β-carotene.

Figure 36 shows the ChemRich enrichment using the top 54 best ranked sampling predictions as opposed to the 54 studied by Thiele *et al.*, for Fish-eye disease/ LCAT deficiency. Indeed, the enriched compound classes here are oligopeptides and retinoids, and not amino acids like previously. Fish-eye disease/ LCAT deficiency, as the name suggests, affects eyesight, resulting in corneal opacifications. Retinoids are class of chemical compounds that are vitamers of vitamin A, which plays a vital role in maintaining a clear cornea [154]. Based on the literature, this appears to be a more coherent enrichment than the previous amino acids enrichment.

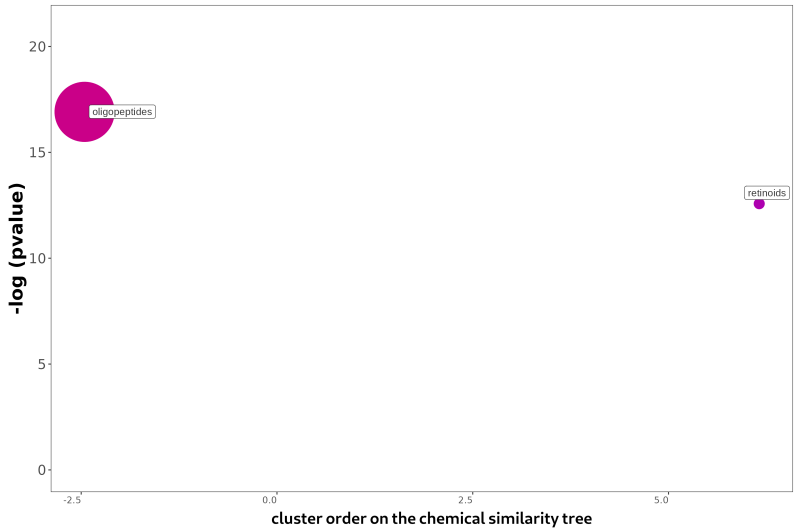


Figure 36: ChemRich using the top 54 metabolite sampling predictions for Fish-eye disease/ LCAT deficiency.

3.5.3 Biochemical distance between altered reactions and predicted metabolites

As expected, some top ranked metabolites are substrates or products of the altered reactions. Nevertheless, we discovered through network analysis that more indirect relationships can be discovered between altered reactions and top ranked metabolites. To do so, the metabolic network was first converted into a bipartite graph using the Met4J library. A bipartite graph applied to metabolism consists of nodes as both metabolites and reactions, but with no edges between any two reactions, and no edges between any two metabolites. All edges cross over from the set of metabolites to the set of reactions, i.e. they alternate between the two sets.

The goal here is to calculate the distance between the reactions that were knocked-out in the SCD condition and the top 50 most changed metabolites according to SAMBA's predictions. The hypothesis is that the closer a metabolite is to the set of knocked-out reactions, the more likely it is to be dysregulated by the disruption. Calculating a distance between two entities using graph theory can rely on the calculation of shortest paths.

In order to calculate distances correctly in a metabolic graph, side compounds must be removed. Side compounds create irrelevant links between reactions when the goal is to calculate how close a reaction is to another. A side compound is usually defined as a metabolite with a high degree of connectivity, i.e. involved in many reactions, and is not considered as the "main" substrate of the reaction. For example, ATP, ADP, and H_2O can be considered as side compounds. Unfortunately, defining a list of side compounds for a network is not an easy task as it can depend on the goal of the analysis as well as the organism or other parameters, and can even be a subjective choice. In this case, a manually curated list of side compounds was used with Human1 to remove edges that would shorten paths too much for the distance calculation to be of use.

The final parameter is choosing whether to use the directed or undirected version of the metabolic graph network. A directed network uses the directionality of each reaction in its catalysed state, whereas an undirected network removes this information when calculating a shortest path. Undirected paths represent the global distance of effect between two entities while directed paths are more indicative of upstream/downstream effects, which is why they were chosen for this study.

Figure 37 shows the distances from the 50 most changed metabolites (rows) as predicted by SAMBA for the SCD condition, in relation to the different reactions that were knocked-out (columns).

Figure 37: Distances from the top 50 most changed metabolites to each of the reactions affected by the SCD SNP. Ranks are shown from beige to dark purple. Distance is shown on the yellow to dark blue-green scale (black is infinite distance, i.e. no path). The distance is measured in the number of reactions and metabolites it takes to get from the KO'd reaction to the extracellular metabolites, using the shortest paths.

This distance heatmap shows that the highest ranked metabolites (top most rows, pale beige annotation on the left) are the closest (yellow) to the perturbed reactions, with distances of 5-6. As we move down the ranked list, the general trend is that the metabolites get further and further from the perturbed reactions, even reaching "infinite" distances, meaning that no path was found between the reactions once the side compounds were removed.

This data can also be represented on a metabolic network visualisation. Figure 38 displays the metabolic subnetwork extracted from Human1 using the perturbed reactions (red squares) and top 50 metabolites (coloured circles).

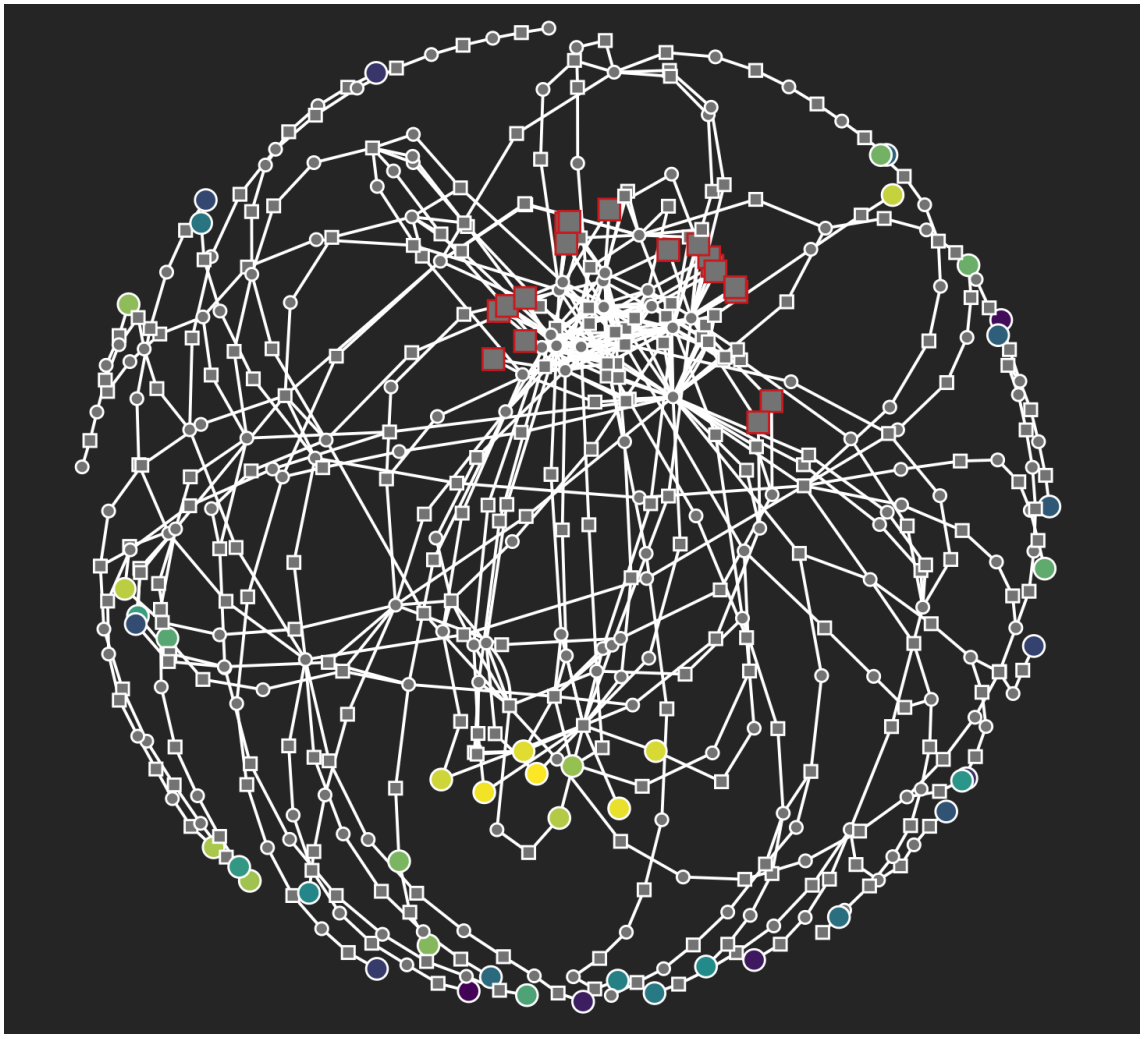


Figure 38: Undirected subnetwork showing paths between the reactions affected by SCD (red squares) and the top 50 predicted most changed metabolites for this condition (coloured circles). Metabolites with a short path to the affected reactions are shown in yellow while darker colours correspond to a longer path (the colour scale is the same as in Figure 37).

Figure 38 shows that while the highly ranked metabolites are relatively close to the KO'd reactions, they are not directly linked to them. Furthermore, many other affected metabolites can be found at a distance of 10 or more steps (shown as periphery metabolites in the circular layout of this figure). These are additional metabolites of interest that could be future paths for analysis which could not

be directly inferred from the affected reactions and scenario. The far but highly ranked metabolites are metabolites we may not have thought of as potential metabolites of interest due to their distance from the disruption in the network, since they may seem unrelated at first glance.

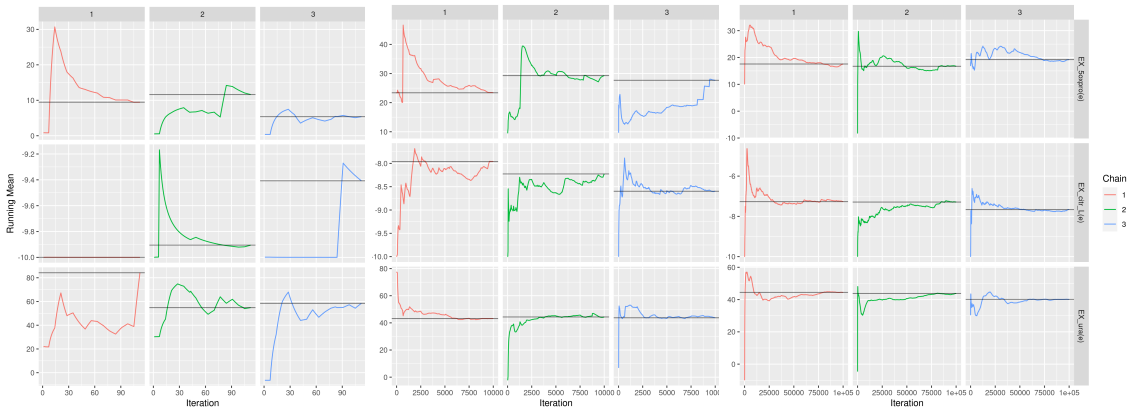
4 Convergence

One of the main limitations of flux simulation with CBM is that it is impossible to fully describe such a large solution space. The method used in this project to explore the solution space is random sampling, but determining a sufficient number of samples is a major challenge when it comes to this method. When using random sampling to explore the solution space, the number of samples to use must be provided, but choosing the ideal number for a given network is a challenge since by definition the structure of the solution space is unknown. The number of samples can be increased, however, in order to sufficiently explore the solution space, a large number (at least 100 000) of samples must be used for larger networks such as Recon2 or Human1 which contain thousands of reactions.

Therefore it is essential to know when to stop sampling: determining when the solution space has been sufficiently sampled. We ran convergence tests using various well-known sampling metrics: running means, traceplots, shrink factor or potential scale reduction factor (PSRF) plots, autocorrelation function (ACF) plots, flux density plots, and partial plots to make sure that using 100 000 samples was enough for a network this large, with the goal of calculating z-scores on distributions. The results can be found in the following figures, where three independent runs for each nsamples value (100, 10 000, and 100 000) are shown for three randomly chosen exchange reactions in the network.

Running means show how the cumulative average value of all sampled fluxes converges with each sampling iteration. If the final value (black horizontal line)

is different between the 3 independent runs (columns per subplot), this shows that there were not enough iterations to reach convergence.

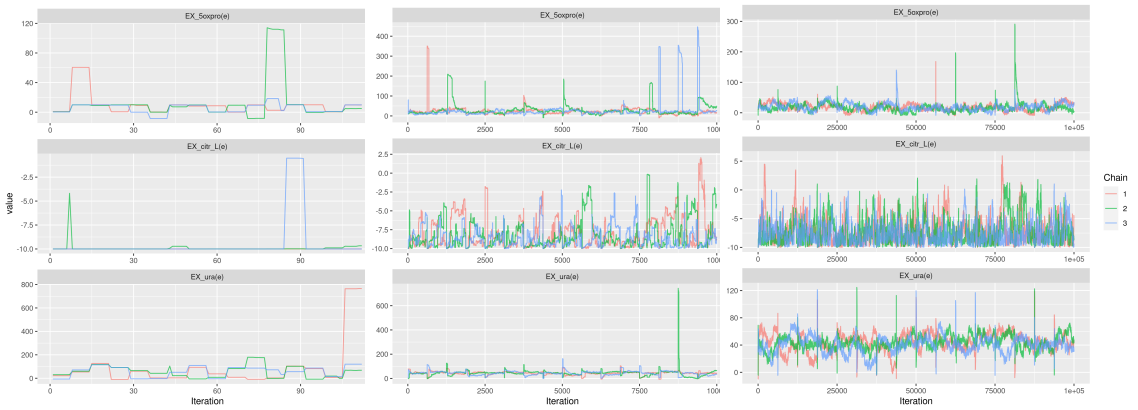


(a) nsamples = 100 (b) nsamples = 10 000 (c) nsamples = 100 000

Figure 39: Running means for 3 random exchange reaction fluxes using (a) 100, (b) 10 000 and (c) 100 000 samples, with 3 independent runs for each. The x-axis shows sampling iterations while the y-axis shows the current mean for that iteration. The three colours are the three independent runs. The rows in each subplot represent the same randomly picked exchange reactions.

The running means clearly show the variability between runs (red, green and blue) when using 100 samples, and still show some variability with 10 000 samples. The running means for 100 000 samples show convergence in each run as well as inter-run stability.

Trace plots show the flux value (y-axis) along iterations (x-axis), for each independent run. Generally, traceplots should show no general trend if there is convergence.

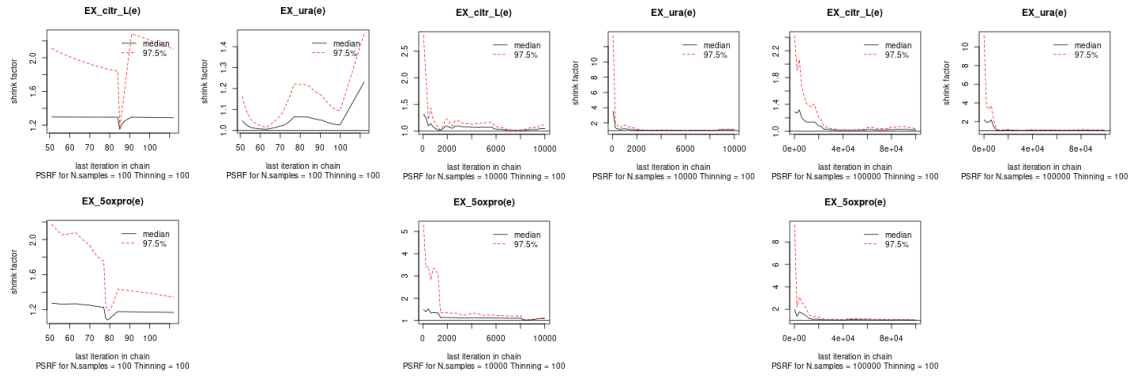


(a) nsamples = 100 (b) nsamples = 10 000 (c) nsamples = 100 000

Figure 40: Trace plots for 3 random exchange reaction fluxes using (a) 100, (b) 10 000 and (c) 100 000 samples, with 3 independent runs for each. The x-axis shows the iterations along the sampling process while the y-axis shows the flux value for each iteration. The three colours are the three independent runs. The rows in each subplot represent the same randomly picked exchange reactions.

For all three nsamples values the trace plots appear relatively stable. For nsamples = 100, the spikes appear larger due to the smaller number of iterations, but these spikes in flux values are normal, even for high nsamples values, as they represent an extreme flux value being found as a solution.

PSRF plots show the shrink factor along iterations. It indicates if runs have "forgotten" their initial flux value. It should decline to 1 as the iterations approach infinity.

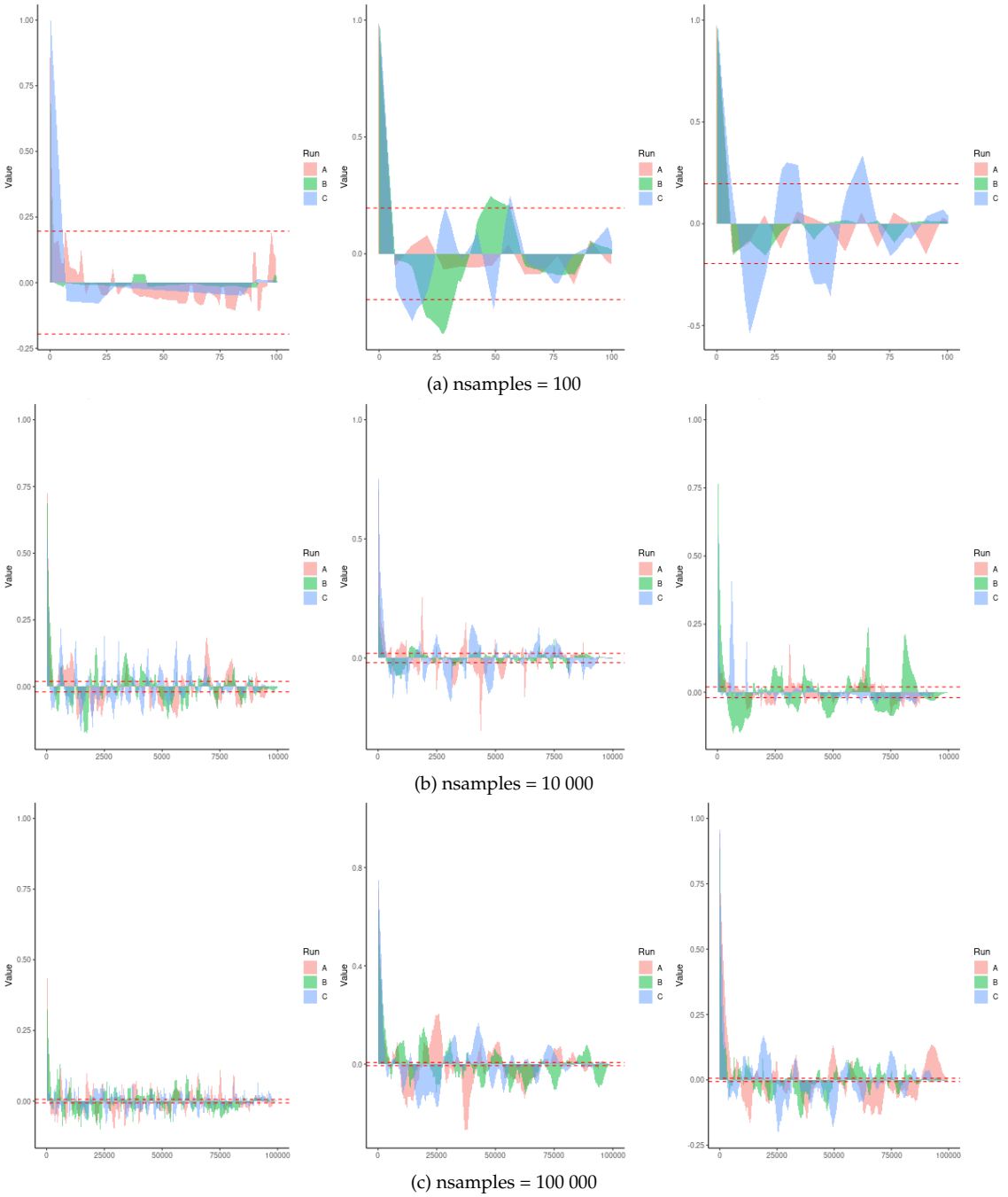


(a) nsamples = 100 (b) nsamples = 10 000 (c) nsamples = 100 000

Figure 41: PSRF plots for 3 random exchange reaction fluxes using (a) 100, (b) 10 000 and (c) 100 000 samples, with 3 independent runs for each. The x-axis shows sampling iterations while the y-axis shows the current shrink factor (PSRF) for that iteration. The three colours are the three independent runs. Each subplot represent the same randomly picked exchange reactions.

The PSRF plots for 100 samples have clearly not converged to 1, whereas those for 10 000 and 100 000 samples rapidly approach 1.

Autocorrelation (ACF) can be measured as a function of the lag along iterations. The sample autocorrelation should decrease as a function of their lag if the chain is properly mixed.



(a) nsamples = 100

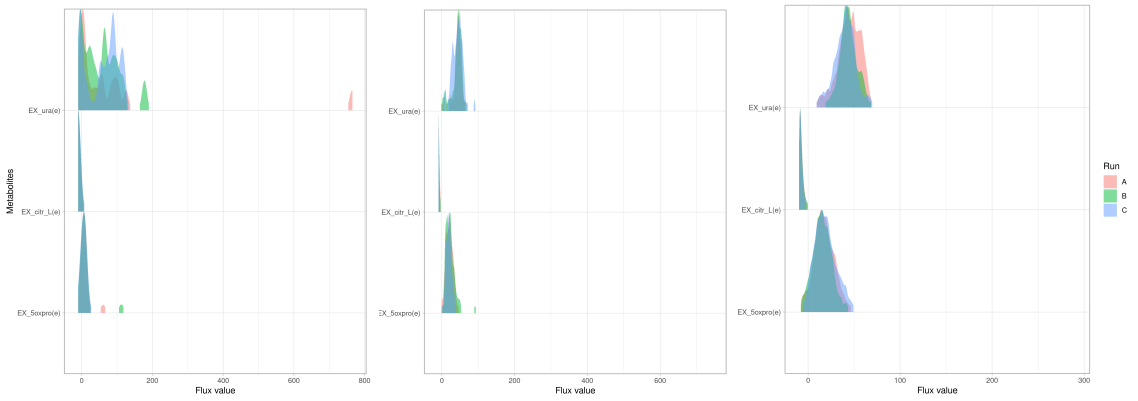
(b) nsamples = 10 000

(c) nsamples = 100 000

Figure 42: ACF plots for 3 random exchange reaction fluxes using (a) 100, (b) 10 000 and (c) 100 000 samples, with 3 independent runs for each. The x-axis shows the iteration lag along the sampling process while the y-axis shows the autocorrelation for each iteration. The three colours are the three independent runs. The columns in each subplot represent the same randomly picked exchange reactions.

The ACF plots show a decrease in autocorrelation relatively rapidly for all three values of nsamples, meaning that regardless of the number of samples, the autocorrelation between iterations is similar.

The flux density plots are a visual way of seeing the smoothness of the sampling distributions across runs, as well as seeing if a given run has managed to explore enough of the solution space. Note that the x-axes in each subplot are not the same scale.

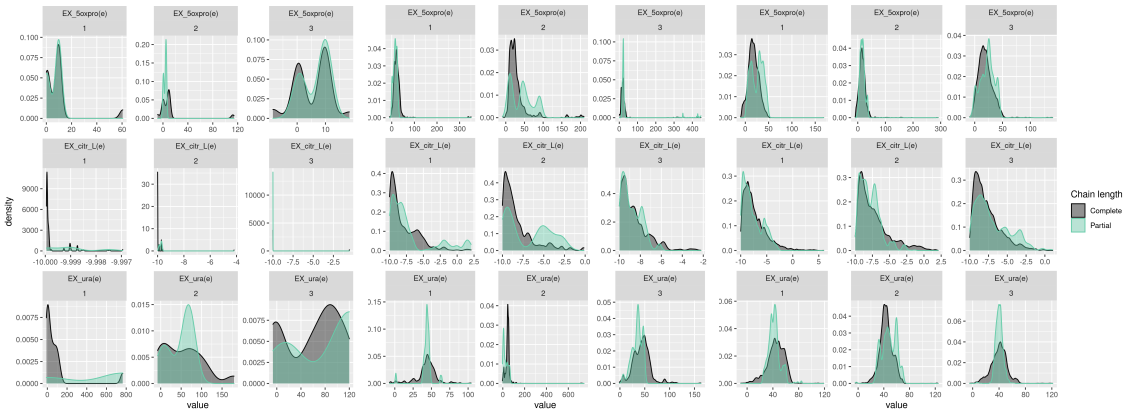


(a) nsamples = 100 (b) nsamples = 10 000 (c) nsamples = 100 000

Figure 43: Flux density plots for 3 random exchange reaction fluxes using (a) 100, (b) 10 000 and (c) 100 000 samples, with 3 independent runs for each. The x-axis shows the flux values obtained at the end of the sampling process while the y-axis shows the density. The three colours are the three independent runs. The rows in each subplot represent the same randomly picked exchange reactions.

When comparing between runs for the same nsamples value, regardless of the value the flux distributions appear relatively stable. However, when comparing between nsamples values (columns), the flux distributions for nsamples = 100, especially the first exchange reaction (top row), are very unstable. Even for 10 000 and 100 000 samples, the first exchange reaction solution space seems more "difficult" to sample for than the other two. We can also see that outliers/extreme flux values with 100 samples are more visible and therefore more weighted, whereas with higher nsamples counts, these values are squashed by the higher flux counts in the center of the distributions.

Partial plots show how the last 10% of samples compare with the whole sampling distribution. This gives a visual preview of if the last samples are representative of the whole distribution. Ideally, the whole and final parts of the chain sample in the same target distribution, so the overlapped distributions should be similar.



(a) nsamples = 100 (b) nsamples = 10 000 (c) nsamples = 100 000

Figure 44: Partial plots for 3 random exchange reaction fluxes using (a) 100, (b) 10 000 and (c) 100 000 samples, with 3 independent runs for each. The x-axis shows the flux values obtained at the end of the sampling process while the y-axis shows the density. The three columns are the three independent runs. The rows in each subplot represent the same randomly picked exchange reactions.

These plots can be difficult to judge visually, as depending on the exchange reaction, the flux distributions can be more or less stable. With 100 samples, the bottom exchange reaction (EX_ura(e)) is clearly undersampled as the final 10% of fluxes (green) do not represent the entire flux distribution (grey) well. With 10 000 samples, again many of the last 10% of distributions do not overlap sufficiently with the complete distributions. With 100 000 samples, there is still a lack of overlap but more distributions overlap correctly.

The main conclusion of this analysis is that one should never rely on one single diagnostic, and that these diagnostic measures cannot guarantee the absence of problems, they only help us to spot a problem. These results indicate that 100 000 samples is sufficient for exploring these large GSMNs, and while not perfect, the flux distributions are enough for computing z-scores for our use case.

Chapter IV

Results Part II: Simulated data for pathway analysis benchmarking

1 Introduction: Pathway enrichment

Pathway enrichment methods were originally developed for analysing genetic expression data. Indeed, with great quantities of data comes great responsibility for analysis and interpretation. As large amounts of data are generated, interpretation relies on reducing the number of variables through approaches like PCA and enrichment analyses to obtain human-comprehensible results that either contain meaning (to us) or from which meaning can be extracted. Gene expression data features can be in the tens of thousands in number and organising this data into smaller functional sets of genes can aid greatly in understanding biological mechanisms.

These smaller sets of genes consist of pathways often using Gene Ontology (GO) [92, 93], from high level sets like ”Immune response” to more fine-grain ”positive regulation of non-canonical NF-κB signal transduction”. They can be metabolic pathways, gene regulation pathways, cell-level responses, cell signalling processes etc.

For metabolites, there are two types of ontologies, both of which are less

widespread than GO is for genes. Chemical ontologies, such as ChEBI, are based on the chemical structure and properties of compounds individually. Compound classes can be ”parents” of multiple metabolites (as shown previously in Figure 32). However, the annotation of compounds in ChEBI is not always optimal, leading to some metabolites being annotated with top-level classes which do not contain much specific information. An example of this is 1-(11Z-icosenoyl)glycerol which is directly annotated as an ”organic molecular entity”. The second type of ontology is biochemical ontologies, which are pathway-based collection of reactions and metabolites as well as sometimes genes. In theory, a pathway ontology would be a good GO equivalent for metabolites, but there is no real consensus and many different pathway ontologies exist, such as KEGG, BioCyc, and Reactome, which all have different pathway definitions. This means that gaining functional information is more difficult and less standardised.

Pathway enrichment analysis serves two primary purposes. The first scenario involves investigating whether specific genes or metabolites of interest are grouped together in a particular pathway. For instance, researchers may want to ascertain whether differentially abundant metabolites resulting from a comparison are produced independently or if they collaboratively participate in shared pathways. In this context, the ORA method proves valuable. The second situation is a more global approach, where the entire list of measured metabolites is considered by establishing a ranking based on a chosen metric. For this, methods like GSEA and MSEA are employed. The rest of this section will refer to pathway enrichment methods in terms of metabolites rather than genes.

1.1 Overrepresentation analysis

ORA involves three key parameters: first, a filtered list of metabolites of interest is required (n in Figure 45). This list can be obtained from an experimental abundance analysis, involving the comparison of metabolite abundance in two conditions. The list must be filtered based on a metric such as fold change using a significance threshold. The second input is a pathway database containing annotated metabolites corresponding to each included pathway (one pathway is shown as k in Figure 45). Finally, an important (but overlooked) input is the background set (N in Figure 45), which is often not explicitly provided by the user, and by default is all metabolites in the network. ORA employs a hypergeometric test on a per-pathway basis to determine the significance of enrichment.

For instance, we can consider a scenario where we have identified 50 (n) significantly abundant metabolites, and our aim is to define how they are connected across the entire KEGG database. Manually, we can begin by taking the first pathway and discerning whether any of the 50 metabolites are present within it. The question then arises: is finding 5 (k) metabolites from our significant list within this pathway considered significant? The null hypothesis for this test is that due to chance alone, more than 5 of our metabolites can be found in this pathway.

To address this question, two factors come into play. First, the size of the pathway is crucial. If the pathway contains 20 (M) metabolites in total, discovering 100% of the metabolites among the initial list of 50 is almost certain to be a significant outcome. However, if our set of metabolites only includes 10 out of 20 metabolites in the pathway, the method needs to consider another aspect: the background set. The background set is defined by the total number of detectable metabolites in the experimental setup. This aspect is pivotal in the analysis, as the significance of finding 10 metabolites in a pathway from a list

of 50 metabolites that have been measured in a targeted metabolomics setup, or 300 metabolites measured by an untargeted setup is not comparable. The background set has a strong impact on the pathway enrichment results and must be taken into account, and in metabolomics this is often overlooked due to the difficulty of defining a background set when many features remain unidentified [43].

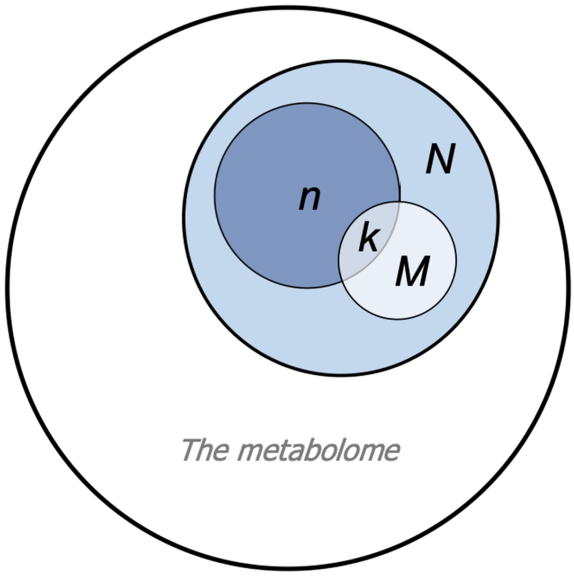


Figure 45: Venn diagram representing ORA parameters. N represents compounds forming the background set, which covers part of the full metabolome. M represents compounds in the pathway of interest. n represents compounds of interest (i.e., differentially abundant metabolites), and k represents the overlap between the list of compounds of interest and compounds in the pathway. Figure from [43].

1.2 Metabolite set enrichment analysis

ORA is often likened to a supervised analysis due to its reliance on a pre-filtered list of metabolites as input. This contrasts with the more global nature of MSEA, which doesn't require a predefined metabolite list. However, MSEA necessitates the availability of as many metabolites in the sample for accurate results, as it takes the entire list of measured metabolites into account. MSEA operates with just two essential parameters: a database of pathways and a sorted list of metabolites. A typical workflow involves ranking metabolites

based on their fold change when comparing two conditions. For each pathway, the MSEA algorithm goes through the sorted metabolite list and it accumulates a score for the pathway. If a metabolite is annotated in the pathway, the score increases, and it decreases otherwise. When multiple annotated metabolites cluster in the sorted list, the pathway's score rises significantly, increasing what is known as the "enrichment score" (ES). To enhance interpretability, this ES is normalized by the size of the pathway, yielding the normalised enrichment score (NES).

Using fold changes or other indicators of increase/decrease as a ranking metric in MSEA can enhance result interpretation. For instance, a significant positive NES implies that the pathway contains increases in metabolite concentrations, indicating an overall "increase" in the pathway's activity, even if this interpretation can be ambiguous regarding metabolites, since an increase in a metabolite can also mean a decrease of the reaction using it as a substrate. The choice of metrics in MSEA is pivotal and can significantly impact the results as well as their interpretation: an alternative approach involves using adjusted p-values from statistical tests to rank metabolites. This alternative metric can provide valuable insights. In this scenario, pathways with the highest NES would be the most deregulated pathways in the comparison, while pathways with a negative NES would represent the most stable pathways. Ultimately, the selection of the appropriate metric and interpretation strategy depends on the research question, the dataset characteristics, and the biological context of the analysis. It is crucial to carefully consider these factors to derive meaningful insights from MSEA results, as with any analysis.

1.3 Limitations and current pitfalls

Pathway enrichment methods have several limits that should be taken into consideration, especially when applying them to metabolomics data. First, there

is always a bias depending on input data used for enrichment analyses: some metabolites cannot be detected using certain setups and will never be able to be used to enrich pathways in these cases.

A major potential pitfall is the definition of each pathway: not only are pathways defined differently depending on the database, but there are issues inherent to how pathways were initially identified and described. A pathway in this case is a set of metabolites involved in a biological function, which is usually the name of the pathway. For example, glycolysis/gluconeogenesis, a KEGG pathway, is split into the glycolysis and gluconeogenesis pathways in HumanCyc. Pathways, by definition, focus on one area of metabolism, usually the metabolism of one or a few metabolites (such as galactose metabolism). These pathways were defined manually, and are also impacted by the order of discovery of metabolites and metabolism throughout history. Only the metabolites the most central to metabolic functions were studied first, and this created a focus on certain metabolites that may not be as central as first imagined. There are several published approaches to reconstruct metabolic pathways automatically, but the results often require manual post-processing expertise [44]. Stoichiometry-based approaches appear to be the most promising but can often lead to thermodynamically infeasible cycles and high computation times [155].

Additionally, metabolite identifier mapping is a major issue: metabolites have multiple names, IDs, isomers and are named differently depending on the database.

Another point to take into consideration is the localisation of measured metabolites. If the metabolomics experiment measures metabolites in the extracellular medium (exo-metabolomics), the possible pool of measurable metabolites is different compared with internal cellular measurements, since not every metabolite is exported out of the cell. Additionally, when focusing on exo-metabolomics, one must consider that much can happen between the origin

of the metabolic perturbation and the export of the metabolite, and biofluid-level metabolites are seen as further away from the internal metabolism than cellular metabolite levels.

Due to its nature, metabolomics data has a lower metabolome coverage than genes have of genome coverage, and there is less metabolite data in general from a single experiment, due to loss of metabolites along the process of experimental setup design, measurement, statistical analysis, and identification.

Despite these drawbacks, pathway enrichment analyses are widely used, often without taking into account the biases caused by the nature of metabolomics data, and therefore treating it as genomic data. Little research has been done on the extent of the effects of the different parameters on pathway enrichment results.

2 Benchmarking pathway enrichment methods using experimental data

The first part of this research study was to benchmark pathway analysis methods using existing metabolomics data, in order to explore the variables and parameters that can change enrichment outcomes. This work was done in collaboration with colleagues from Imperial College London, and was published in 2021 [43]. The first author of this paper, Cecilia Wieder, carried out the main tasks consisting of varying the different inputs of ORA to determine those with minimal to extreme impacts on the results, described in the following paragraphs. Various public MS datasets were used for analysis due to their non-targeted and multi-species nature.

The first ORA parameter that was analysed for this paper was the background set. Indeed, intuitively, the background set has a large impact on the significance of results but until this study, this was not quantified. It is important to only

include metabolites that are measurable in the sample, in the background set, as including a generic background set means that extra metabolites are taken into account in the statistical test. The results showed, by varying the background set from non-specific to specific, that not only were more pathways significant when using the non-specific set, but also certain pathways were significant in one case but not in the other. For some datasets, the only way to get significant pathways when using a specific background set was to increase the p-value cut-off to 0.1 for demonstration purposes (Figure 46). These differences are mainly due to the size differential between the sets, with smaller background sets causing less pathways to be enriched.

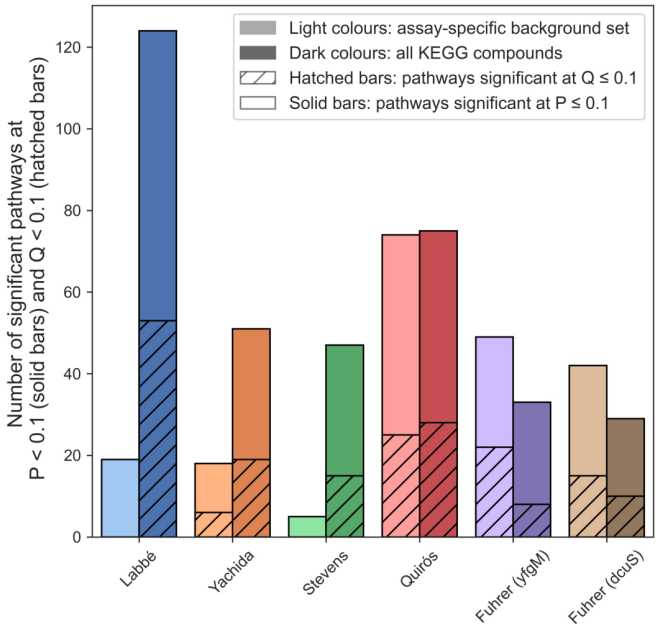


Figure 46: Number of pathways significant at $p \leq 0.1$ (solid bars) and the number of pathways significant at $q \geq 0.1$ (hashed bars, BH FDR correction). Datasets are ordered by number of compounds mapping to KEGG pathways. Figure from [43].

The second input is the list of differentially abundant metabolites. The chosen threshold is a major parameter in defining this list and can have a large impact on results, since any metabolites under the threshold are not taken into account for enrichment. When increasing the threshold, the metabolite list increases in size leading to more significantly enriched pathways up to a certain point, eventually reaching the entire background set list, resulting in zero significantly enriched

pathways (shown in Figure 47).

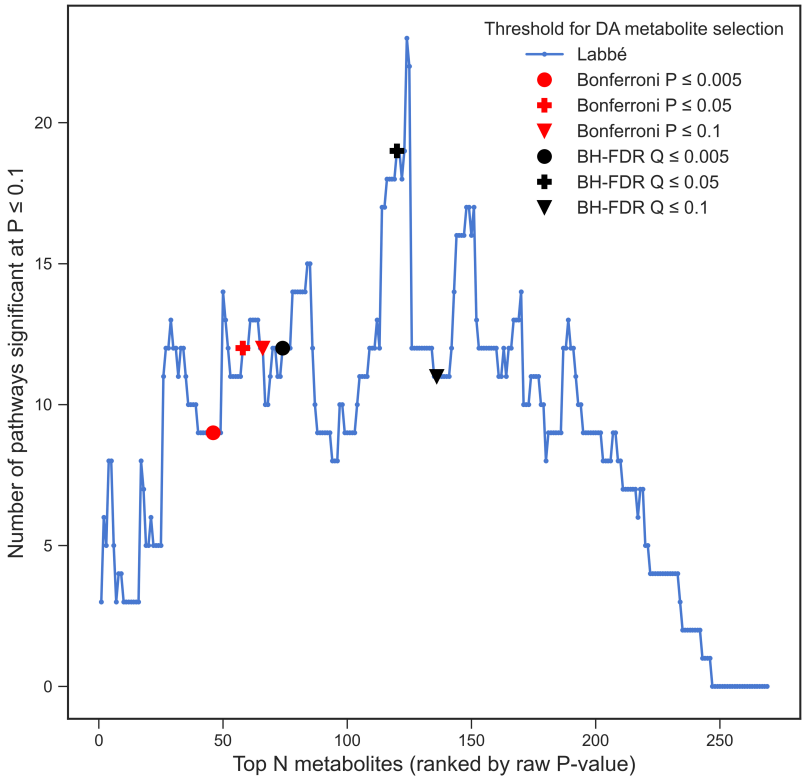


Figure 47: The effect of the number of DA metabolites in the list of metabolites of interest on the number of significant pathways ($p \leq 0.1$) in the Labbé et al. dataset. Results corresponding to Bonferroni thresholds are denoted by red markers while those corresponding to BH FDR thresholds are denoted by black markers. Marker shape (circle, cross, or triangle) represents the adjusted p-value threshold for DA metabolite selection (0.005, 0.05, and 0.1 respectively). Figure from [43].

Another major input in ORA is the pathway database used to define pathways of the metabolic profile. Pathway sets can vary in size and nature between databases. KEGG, Reactome and HumanCyc were tested and the enriched pathways were compared between each test, resulting in a low concordance of pathway names for the significantly enriched pathways. The overlap of metabolites between the significant pathways was also low, as quantified by an overlap coefficient.

Finally, variations on these inputs, simulating experimental issues, were tested, such as the effect of metabolite misidentification, as well as chemical biases induced by the experimental setup. This resulted in certain significant

pathways which were enriched due to misidentified metabolites, as well as loss of previously significant pathways. The chemical bias analysis revealed that certain areas of metabolism cannot be accessed due to the differences in experimental setups.

This work concludes by providing the community with guidelines to avoid the misuse of pathway enrichment methods for metabolomics data analysis, as well as the information that should be reported when performing any type of pathway enrichment analysis. The recommended guidelines are the following:

- Use a realistic background set specific to the experimental setup, usually the entire set of identified metabolites.
- Use an organism-specific pathway set if available, and perform enrichment multiple times with different pathway set databases to perform a consensus enrichment.
- Use multiple testing correction for differential metabolite selection as well as significantly enriched pathways if possible.

This work contributes greatly to our understanding of these methods that were originally developed for gene expression data, but are now used without hesitation on other, less adapted types of data. The missing information to push this benchmark to the fullest is knowing the metabolic disruption causing the metabolic profile. To gain better insight on the mechanics of metabolism and how they are revealed using pathway enrichment methods, the underlying disruption must be known, which will enable the identification of true positive enrichments. This is explored in the following section.

3 Benchmarking pathway enrichment methods using simulated data

In order to fully benchmark pathway analysis methods in metabolomics, the "true positive" state must be known. Indeed, we need to be able to match the enriched pathways with the pathways that were actually disrupted in the condition. The main problem is that this sort of experimental dataset does not exist for humans. By forcing the disruption ourselves using simulations, we can know exactly where the metabolic disruption occurred.

The goal here was to benchmark pathway enrichment methods using SAMBA to create metabolic disruptions in known pathways, attempt to obtain these pathways as enriched (Figure 48), and perhaps show the flaws in pathway enrichment methods when applied to metabolic data. The work in this section is still in its preliminary stages but serves as a first exploratory step into testing these methods.

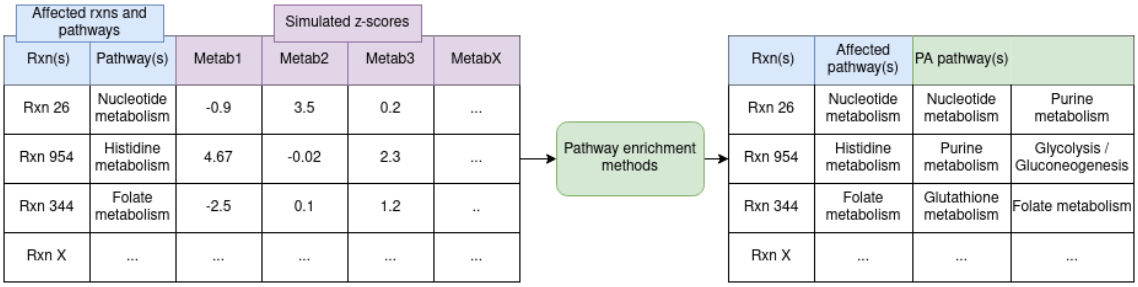


Figure 48: Benchmarking pathway enrichment methods using simulated data. Affected reactions and pathways in the network are in the blue columns. Simulated metabolite fluctuations are in purple, and enriched pathways using these simulated metabolites are in green.

By affecting a reaction in a given pathway (blue columns) in a metabolic network followed by predicting a metabolic profile for that disruption (purple columns), we can expect to obtain the disrupted pathway in the top most enriched pathways using traditional pathway enrichment methods (green columns).

Additionally, by using predicted metabolic profiles, the exact background set

for enrichment approaches is known: the entire list of predictable metabolites. The pathway sets are those used in the metabolic model which means there is no information loss due to identifier mapping.

In this section, SAMBA was run on Human1 by blocking all of the reactions in each pathway independently, and predicting the metabolic profile for each blocked pathway condition. This creates a disruption condition of an entirely blocked pathway for each pathway in the network, only if this blocking this pathway does not stop growth. If the entire pathway KO renders the model non-viable, the flux simulations cannot take place and therefore a metabolic profile cannot be predicted for that condition.

3.1 ORA and MSEA enrichments

3.1.1 Preliminary enrichment

ORA and MSEA were run using the metabolic profiles predicted for each total pathway knockout. For ORA, metabolites were filtered by calculating a p-value from the z-scores and keeping those strictly below 0.05. The input for MSEA was created by ranking all metabolites by their z-score. Figure 49 and Figure 50 show the average confusion matrices for ORA and MSEA with true positives (TPs), true negatives (TNs), false positives (FPs) and false negatives (FNs). They are calculated by taking the average of all predictions across all conditions. If a pathway KO produces a metabolic profile that then enriches that same pathway significantly using ORA or MSEA, it is counted as a true positive hit. A false positive pathway is a pathway that is significantly enriched despite it not being knocked-out in the flux simulation. A false negative corresponds to not finding the knocked-out pathway as significantly enriched.

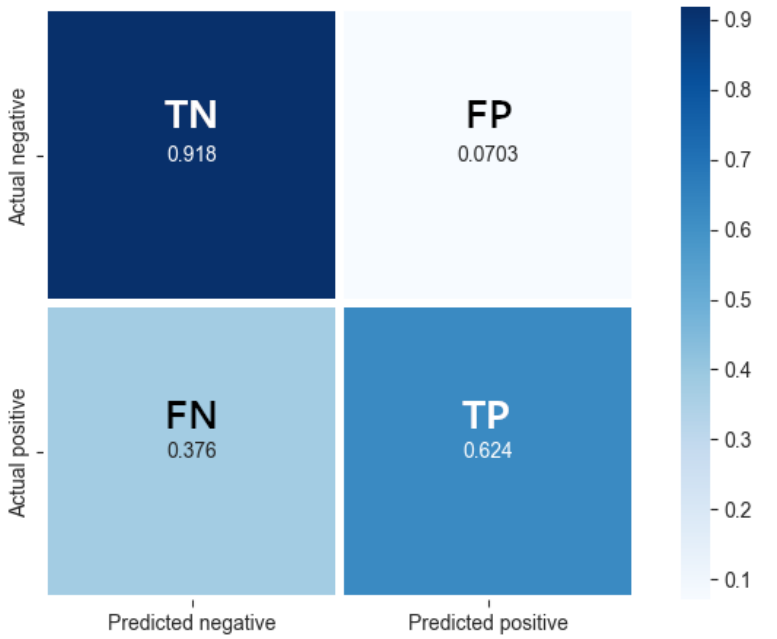


Figure 49: Average confusion matrix for ORA.

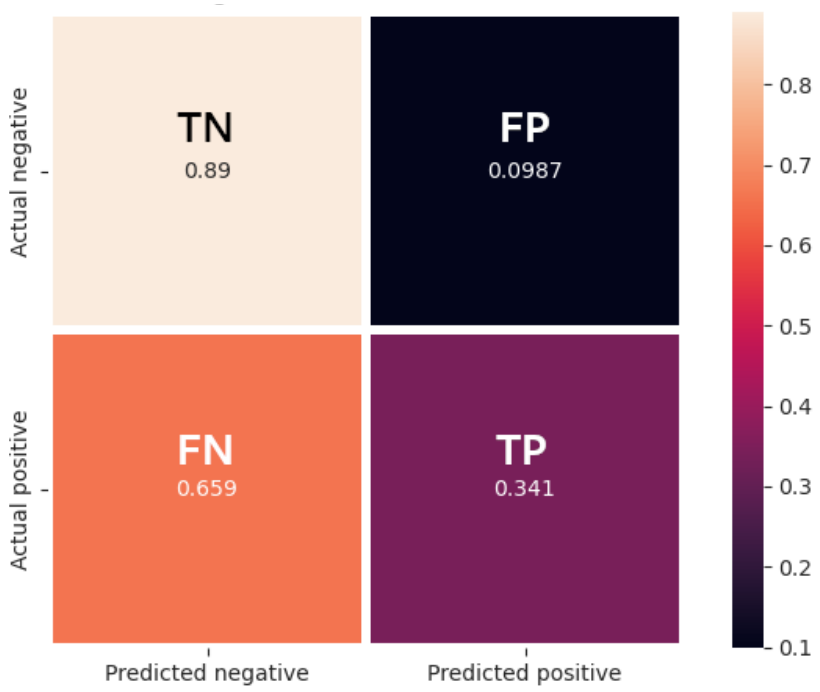


Figure 50: Average confusion matrix for MSEA.

Confusion matrices are generally good indicators of prediction performances for methods like this. In this case, they show the balance between sensitivity (capturing as many TPs as possible) at the expense of specificity (higher FN rate).

Despite the high number of FNs, in both cases, the matrices show that when a positive is predicted, it is very unlikely to be wrong, even if we are not able to capture all positives (37% and 66% of missed positives). The differences between the ORA and MSEA predictions are explored in Section 3.1.3.

The high number of FNs, corresponding to pathways that were not significantly enriched despite them being knocked out in the model, is intriguing as it could reveal the issues with pathway enrichment methods for metabolomics data, especially with enrichments possibly being unrelated to the original perturbation. This can confirm the initial motivation behind this work: by using metabolites as input for pathway enrichment, if these metabolites are indirectly linked to the disrupted pathway, the enrichment may return incorrect results.

3.1.2 False positive example

In order to investigate the reasons behind false positive hits, an example is detailed below. The pathway that was entirely knocked-out in the network for the MUT condition is Acylglycerides metabolism. This produced a metabolic profile of extracellular metabolites, which was then filtered by keeping the top 25% of z-scores. The list of significantly changed metabolites was then used in ORA and resulted in the following significantly enriched pathways, shown in Table 8. The knocked out pathway is shown in red, while other false positives are shown in blue.

ID	Hits	Coverage	P-value	P-adjust
Acylglycerides metabolism	5/18	18/19	6.19E-08	1.11E-06
Glycerophospholipid metabolism	4/51	51/55	0.000335	0.002012
Glycerolipid metabolism	3/25	25/27	0.000628	0.002828
Glycosylphosphatidylinositol (GPI)-anchor biosynthesis	2/14	14/14	0.004287	0.015433
Prostaglandin biosynthesis	2/41	41/41	0.034645	0.102045

Table 8: Significantly enriched pathways after running ORA on the simulated metabolic profile predicted with SAMBA using the total pathway knockout of Acyglycerides Metabolism. The true positive result is shown in red, and false postives are shown in blue.

In this case, the knocked-out pathway is enriched significantly and is even the most enriched when comparing hit ratios ($\frac{5}{18} = 0.28$ vs $\frac{4}{51} = 0.08$), and is therefore considered as a true positive hit. The four other pathways are considered as false positives as they were not affected in the MUT state of the flux simulation, but these extra pathways could also be new information indicating the effects of the original pathway KO.

The following Figure 51 shows a network view of the pathways listed in Table 8. The pathway highlighted in red (edges) is Acyglycerides metabolism. Other pathways from the table are shown in varying colours. Because the metabolic profile consists of extracellular metabolites, in order to map them onto the network which in this case consists of internal pathways, they must be converted into their corresponding intracellular versions. For instance, an extracellular metabolite M_e may also exist in the cytoplasm and the mitochondria, which means we can map the M_c and M_m versions of the metabolite onto the network. The metabolites shown in red are all of the compartment variations of the significantly enriched metabolites for this condition.

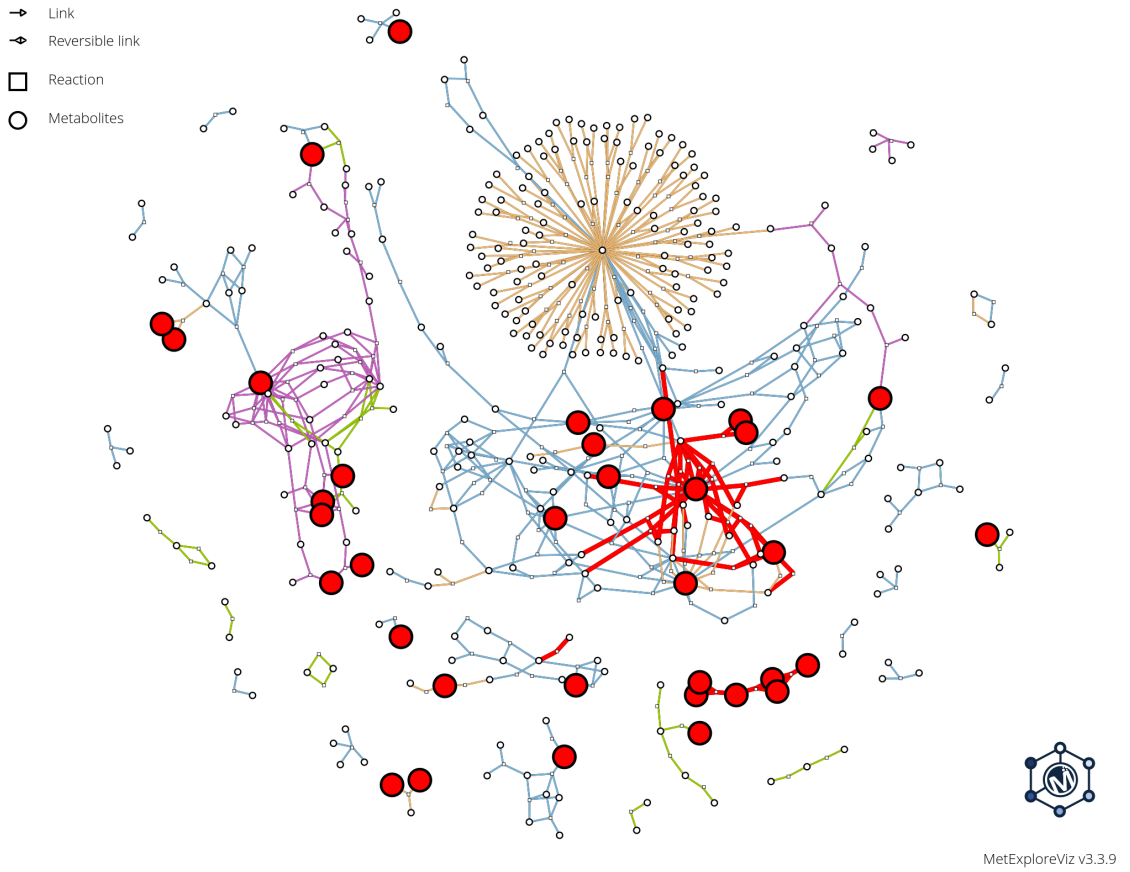


Figure 51: Network view of the blocked pathway, the simulated metabolic profile, and other significantly enriched pathways. The acyglycerides metabolism pathway is shown in red. The other significantly enriched pathways (false positives) are shown in other colours. The metabolic profile used to enrich these pathways is shown as red circles.

The network view shows that metabolites not directly within the knocked-out pathway are affected despite the distance between them. This causes other pathways to be enriched which may or may not be relevant to the KO condition.

3.1.3 Using absolute z-score values

In the previous section, MSEA had a lower true positive rate than ORA which can be surprising. In order to understand this difference, the use of absolute values of z-scores instead of raw z-scores was investigated. However, one must note that by using absolute values, the directionality of the z-scores is lost. The highest ranked absolute z-scores represent the most deregulated metabolites,

regardless of their increase or decrease. Consequently, the MSEA NES must be interpreted differently: the NES in this case will signify strong to weak enrichment, as opposed to a directional overall pathway increase or decrease.

Figure 52 shows a standard MSEA plot of an enriched pathway. Due to the bidirectionality of z-scores (as shown in Figure 53), significant metabolites are spread between the extremely high ranks (red) and extremely low ranks (blue).

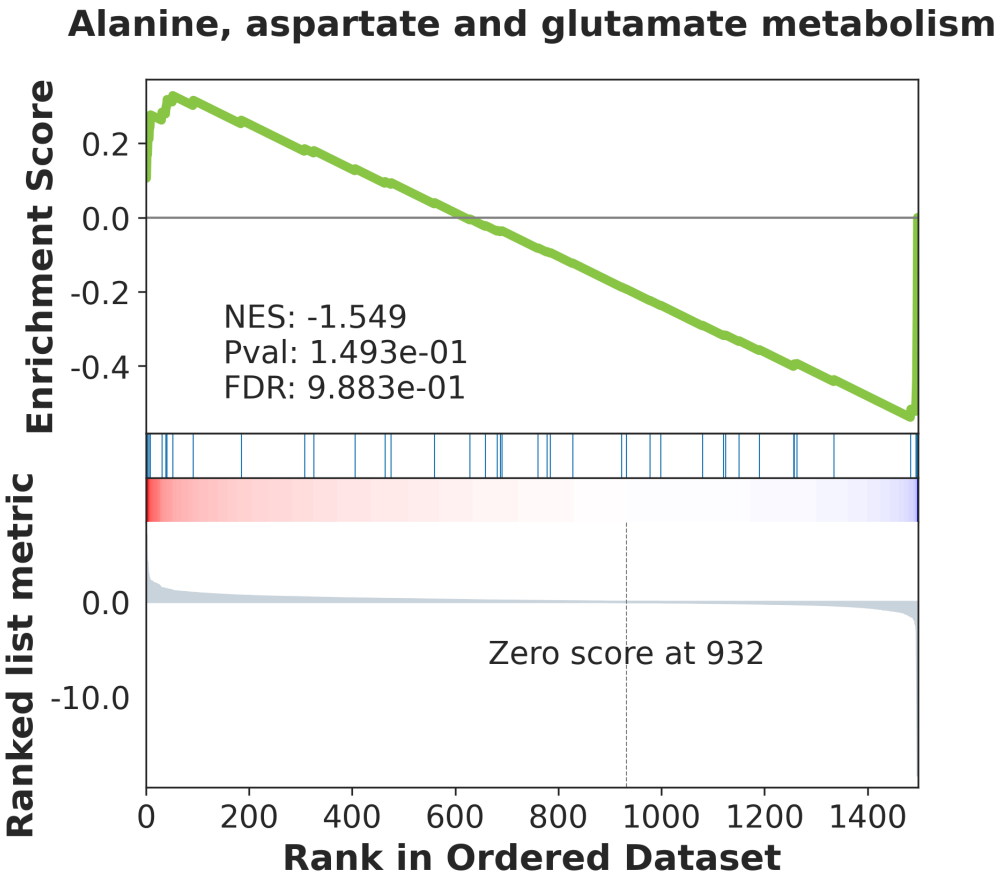


Figure 52: Example MSEA plot.

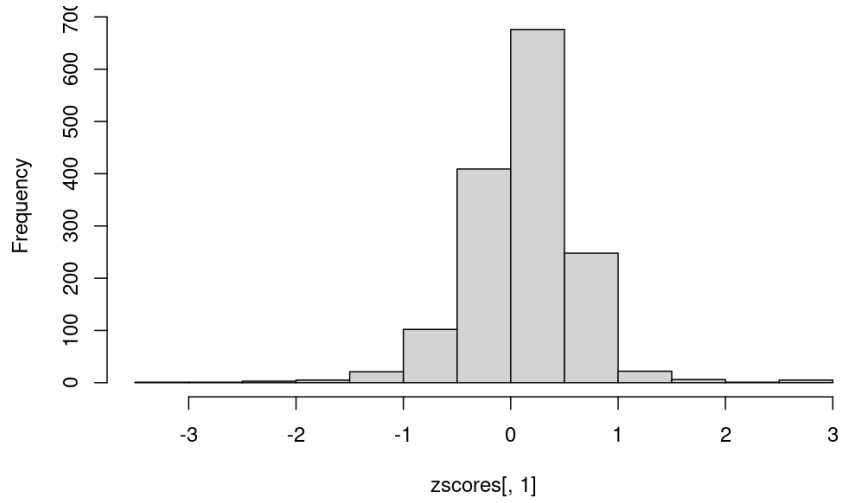


Figure 53: Example z-score distribution of a simulated metabolic profile.

By converting the z-scores to absolute values of the z-scores, the following distribution can be observed in Figure 54.

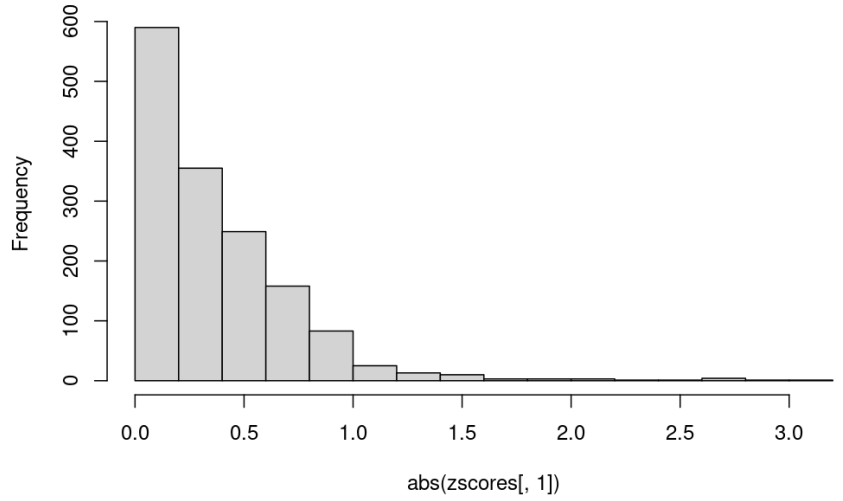


Figure 54: Example z-score distribution of a simulated metabolic profile using absolute values.

The average MSEA confusion matrix when using the absolute values of z-scores to rank the input metabolites is shown in Figure 55. The true positive rate increased from 0.341 to 0.635 when using absolute z-score values instead of raw z-scores.

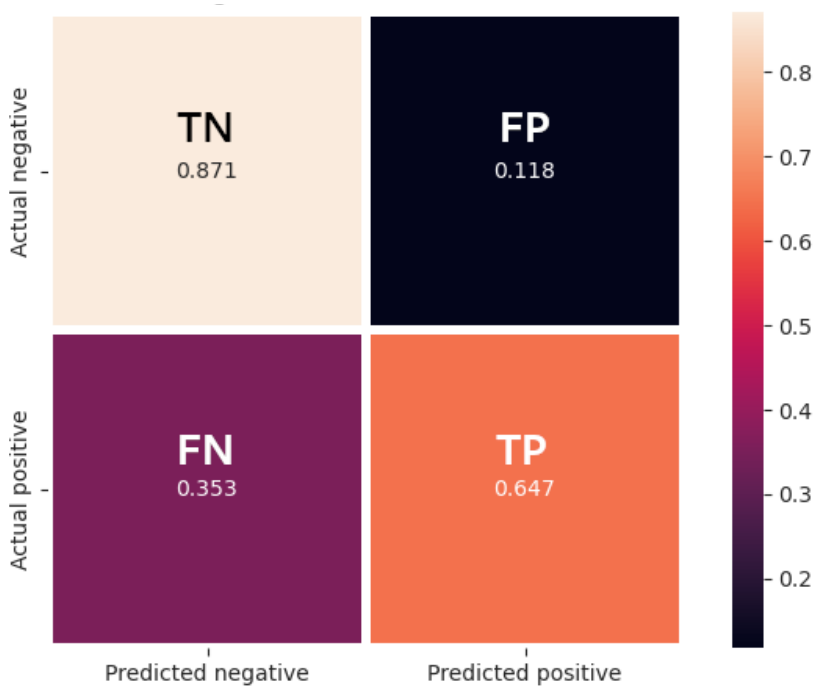


Figure 55: Average confusion matrix for MSEA when using z-score absolute values to rank the MSEA input list.

Normally, MSEA is able to take into account directionality when calculating enrichment for a list of ranked entities ranging from negative to positive scores. In this case, grouping the most differentially abundant metabolites together results in better predictions regardless of the directionality of change. This is visible in Figure 56 where most of the significant metabolites are shown on the left of the plot (and are highly ranked), increasing the NES drastically.

This difference could be explained by the fact that an increase in metabolite exports does not mean the corresponding pathway is necessarily upregulated. As a hypothetical example, in "Alanine, aspartate and glutamate metabolism", alanine could be increased while aspartate is decreased, leading to a split enrichment which is ignored when only using the fact that both alanine and aspartate are significantly differentially abundant, regardless of direction. This is in contrast with gene enrichment where often sets of genes are upregulated together, leading to a general pathway upregulation.

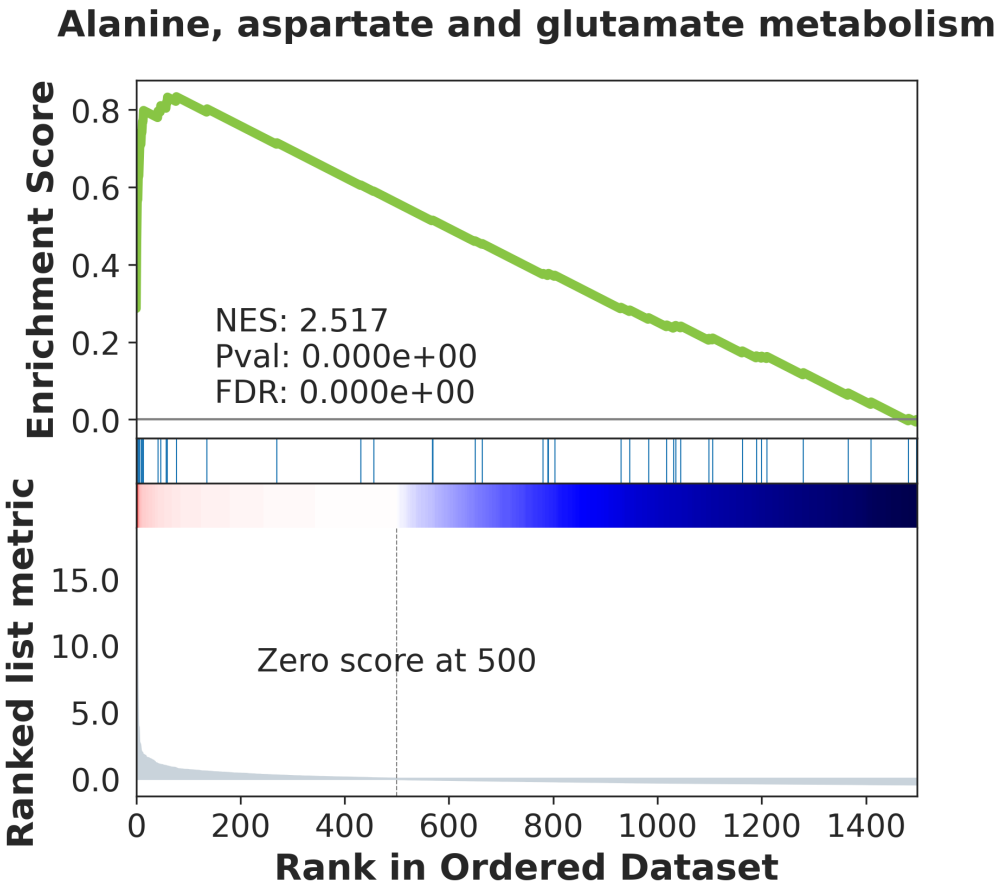


Figure 56: MSEA plot using absolute z-score values as input.

3.2 Distance analyses using graphs

The hypothesis for the following section is that a metabolite that is highly ranked using SAMBA on an entire pathway knockout is close in network distance to that pathway. For this, I computed shortest paths between the metabolites involved in the knocked-out pathway and the most differential metabolites (ranked by absolute value of z-scores), for each pathway simulation. The distance is measured in the number of metabolites it takes to get from the KO'd pathway metabolites to the extracellular metabolites, using the shortest paths in the undirected network. Figure 57 and Figure 58 show two examples of heatmaps representing these distances, using the top 50 most changed metabolites (rows) for each pathway KO metabolites (columns).

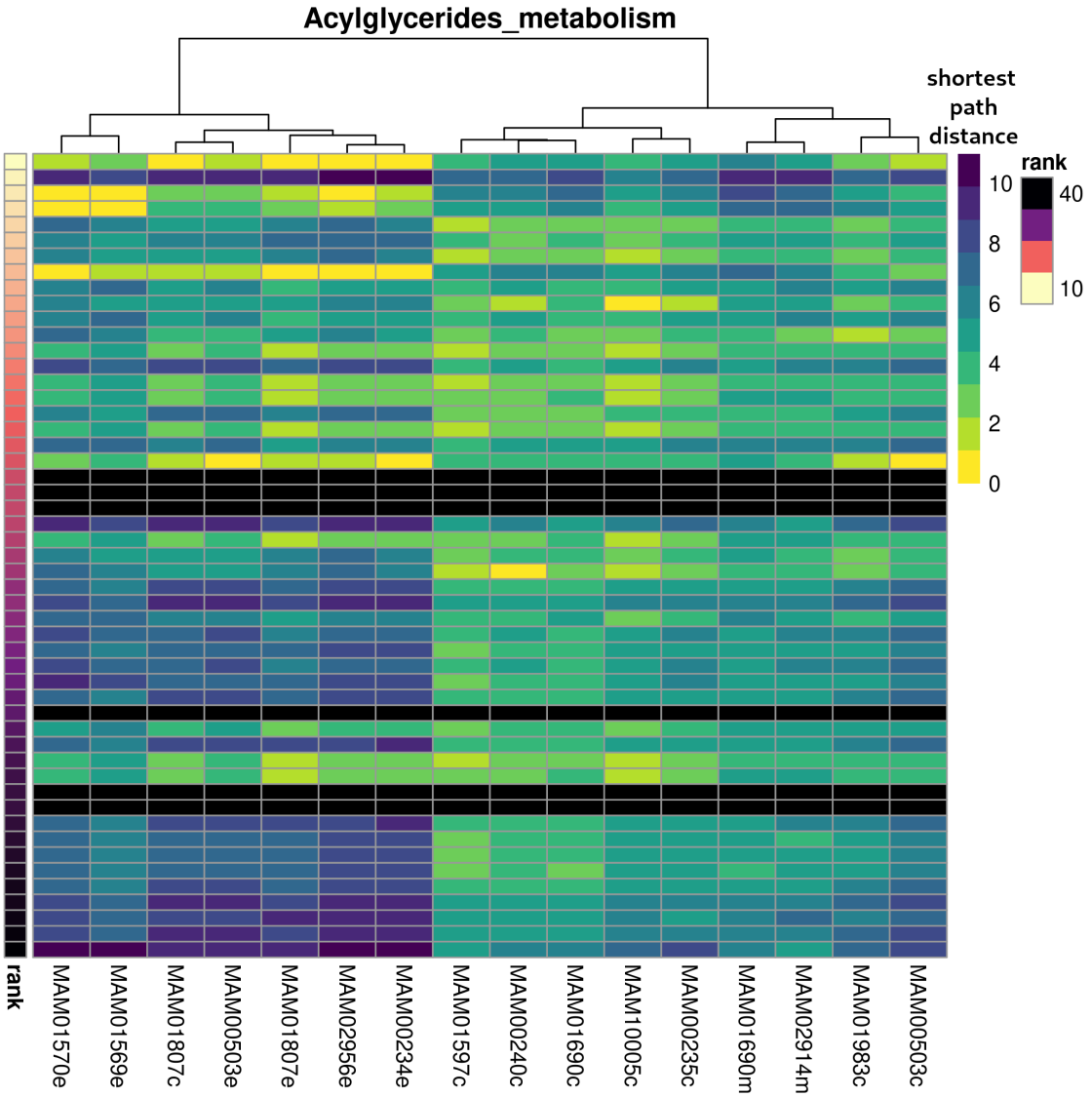


Figure 57: Distances from the top 50 most changed metabolites to each of the metabolites in the KO'd pathway Acylglycerides metabolism. Ranks are shown from beige to dark purple (left column). Distance is shown on the yellow to dark blue-green scale (black is infinite distance, i.e. no path). The distance is measured in the number of metabolites it takes to get from the KO'd pathway metabolites to the extracellular metabolites, using the shortest paths in the undirected network.

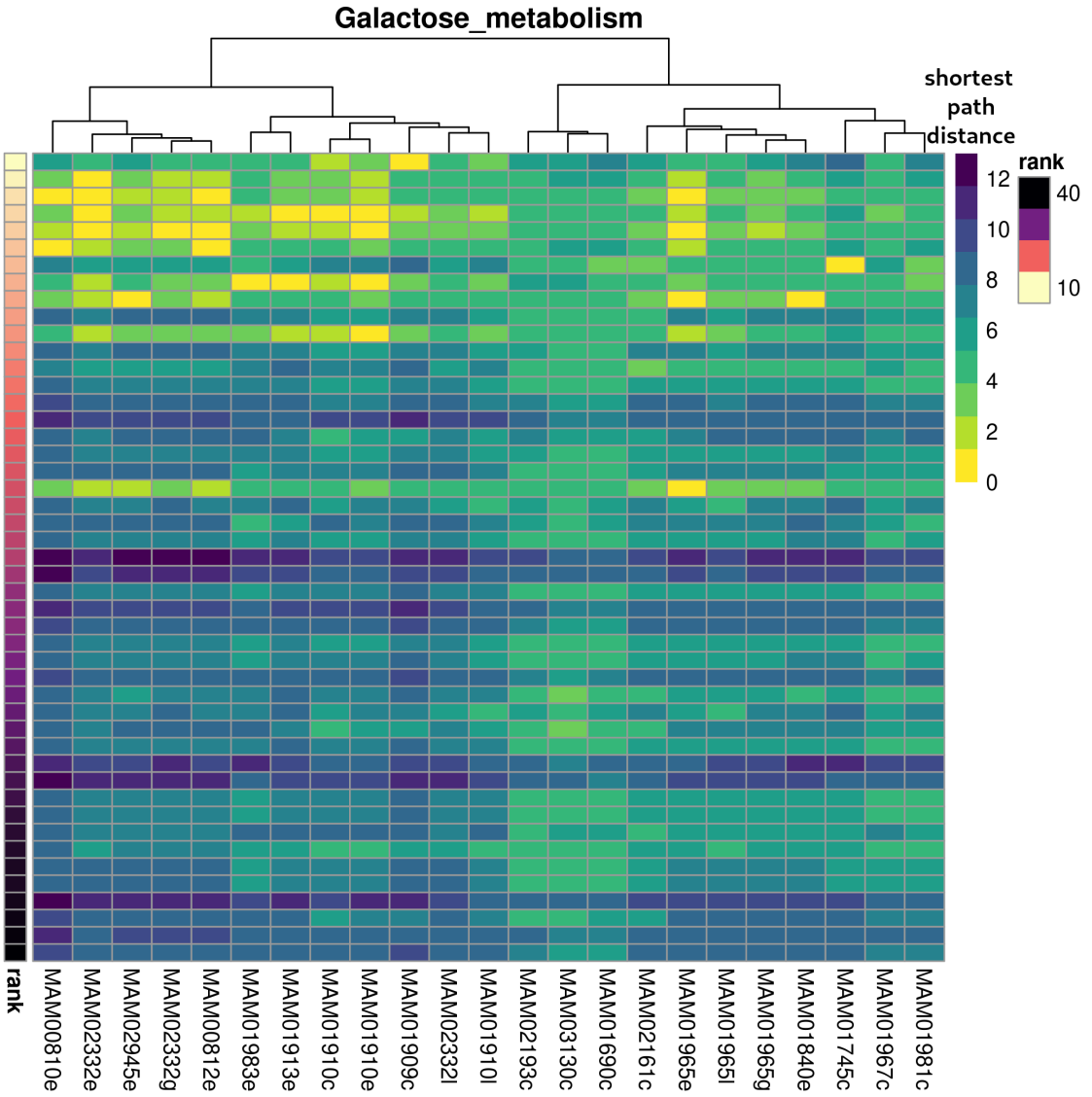


Figure 58: Distances from the top 50 most changed metabolites to each of the metabolites in the KO'd pathway Galactose metabolism. Ranks are shown from beige to dark purple (left column). Distance is shown on the yellow to dark blue-green scale (black is infinite distance, i.e. no path). The distance is measured in the number of metabolites it takes to get from the KO'd pathway metabolites to the extracellular metabolites, using the shortest paths in the undirected network.

The hypothesis for these analyses is that the further down the ranked list we go, the greater the distance to the knocked-out pathway metabolites. This trend is somewhat visible on some heatmaps (Figure 58, left cluster of Figure 57) but was not generalisable to all pathway KOs. Some distances are at 0 because the top ranked metabolite is the metabolite from the pathway. For example, the third best ranked metabolite (3rd row) in Figure 57 is at distance 0 of the first

pathway metabolite (1st column) because it is MAM01570e, and is therefore in the pathway. Figure 59 shows a subset of Figure 57 with only the rows (ranked metabolites) that are involved directly in the Acylglycerides metabolism pathway.

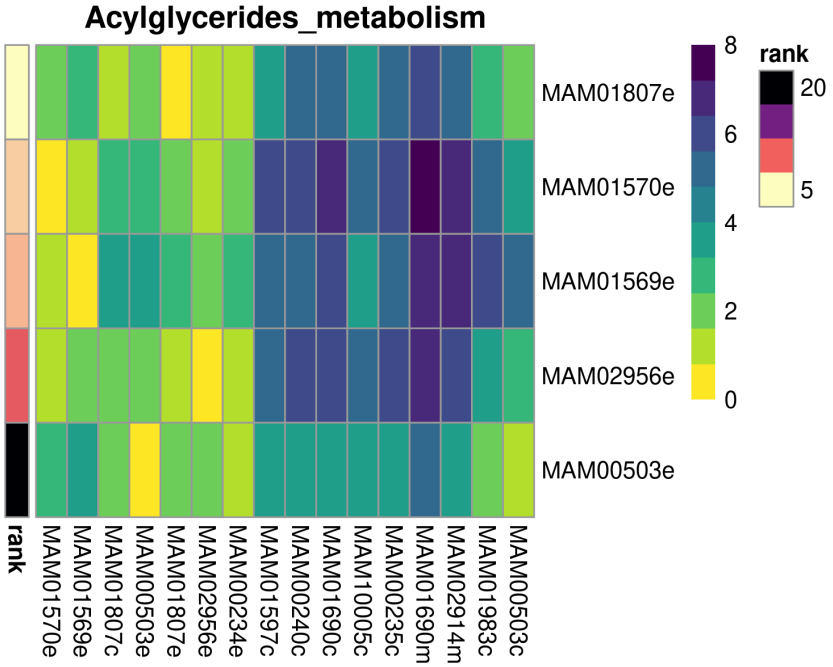


Figure 59: Subset of Figure 57. Distances from the 5 metabolites involved directly in Acylglycerides metabolism to each of the metabolites in the KO'd pathway Acylglycerides metabolism. Ranks are shown from beige to dark purple (left column). Distance is shown on the yellow to dark blue-green scale (black is infinite distance, i.e. no path). The distance is measured in the number of metabolites it takes to get from the KO'd pathway metabolites to the extracellular metabolites, using the shortest paths in the undirected network.

4 Conclusion

To conclude on this chapter, pathway enrichment analyses appear to be less adapted to metabolomics data due to experimental factors as well as the fundamental nature of how metabolites themselves are linked biologically. When there is a metabolic perturbation, the disrupted metabolites can quickly be far in terms of number of reactions from the origin point. The work on benchmarking using real data displayed the limits of pathway enrichment for

metabolomics data. The results shown using simulated data with pathway enrichment investigated the question of distance between the disruption site and circulating metabolites, a step towards quantifying the risk of errors when employing this method.

The next steps to push this further are to link the graph-based distances back to the enrichement results, for example by using the lead metabolites list that MSEA provides to see if they correspond to the closest metabolites to the disruption or not. The graph and metabolic network can also be used be used to our advantage to provide visual information on where these lead metabolites are located in relation to predicted metabolites and the disruption. More perturbation scenarios can also be tested, with different combinations of pathway knockouts, as well as flux reduction instead of knockouts. This work is the first step to creating a simulated benchmark for future development of new metabolomics-centric pathway enrichment methods.

Chapter V

Discussion, conclusion and perspectives

1 Discussion

The results presented in this thesis show that by using metabolism-simulating methods like SAMBA, entire metabolic profiles can be predicted. For instance, for the SCD case study, the metabolites reported as associated with the SNP were highly ranked, especially when considering the total number of exchange metabolites in the whole human network. By adding information from the non observed but highly ranked predictions, we were able to extend the experimental metabolic profile to include new potential metabolites of interest for future studies.

1.1 Technical limitations and scoring discussion

We demonstrated that sampling can add a layer of information to better improve metabolic profiling compared to FVA. Sampling provides a finer grained description of changes which helps order metabolites based on their likelihood to be affected by a perturbation. Compared with FVA, sampling is more computationally intensive (CPU and memory) but recent strategies are

reducing this computational burden [156, 157, 158]. Nevertheless, sampling is currently more than feasible on large networks such as Human1.

The samples ACHR-based samplers generate are not guaranteed to converge to a uniform distribution due to the dependence of each sample on the previous sample. This means that the number of samples one must define is a major contributor to how converged the resulting samples are. When exploring a large solution space, many more samples are required to ensure we get as close to independence and convergence as possible. This can be improved using the thinning parameter T, which increases the total number of sampled values. The resulting number of samples is the same, but each final sample is distanced by T in the chain of samples, while the rest are discarded. This therefore requires Ns * T total samples calculated throughout the sampling process to result in Ns samples, which can inflate sampling run durations.

The FVA method used in previous work [128] to compare intervals calculates the greatest change between the two pairs of upper bounds and the two pairs of lower bounds. This comparison of boundary shifts is not always representative of the underlying changes and can mislead the interpretation of the intensity of these changes. Using other methods such as comparing the means of boundaries assumes a uniform flux distribution within these bounds, which we have shown via sampling is rarely the case. Using the most frequent fluxes with sampling appears to be a good approximation of the mix of metabolite exports that occurs in biofluids, but it should be noted that the most frequent flux value may not be the most frequently observed flux in reality. However, in some cases, the most frequently predicted flux value may not represent the biological reality of a cell, such as for cells in extreme conditions or fast-growing cancerous cells, for which fluxes might be more close to the extremes. To represent these extreme conditions in SAMBA, the initial parameters of the model could be adjusted (higher minimal production of biomass) to force the model to operate within extreme (boundary) optimums as opposed to more likely fluxes.

The boundary shifts evaluated by FVA are very sensitive to change, since a very low threshold ($1e^{-6}$ tolerance and 0.01 factor) for change is used to report an increase or a decrease. Despite this, FVA is able to predict biomarkers, as shown in previous studies [128, 83], when aiming to predict specific biomarkers. We progressed from the calculation of a score to the ranking of these scores since ranking the change intensities via sampling means that the most changed metabolites can be highlighted, while still keeping information on the other subtle metabolite changes. Contrary to the binary change/no change method of reporting FVA results, sampling ranks provide information on a wider scale by taking into account relative changes between metabolites.

Z-scores prove to be useful in that they reflect an intensity of change similar to fold changes, and are weighted by the standard deviation of the distributions, which helps the z-scores to remain flexible given the variable nature of these distributions. Initially, instead of using a z-score to compare sampling distributions, more widely used statistical metrics were tested, such as Kullback-Leibler Divergence, Kolmogorov-Smirnov, and Wasserstein. However, they did not prove to be informative in our use case since they lead to p-values being too sensitive, resulting in extremely significant p-values for very similar distributions. In addition to this, these tests provide scoring metrics which are unable to quantify or describe the differences in the way a z-score can. Z-scores efficiently capture both the intensity and extent of variation of flux distributions between conditions. Additionally, we assessed various other metrics in order to decipher their ability to capture relevant metabolite rankings (Figures 12 and 13).

The comparison of probability distributions is not a problem specific to random sampling: distributions are used in different fields such as statistics. The simplest way to compare two distributions is to calculate the mean or median of each distribution and compare those. This however follows the assumption that the distributions are normally distributed, which in general is not always the case. There are more complicated statistical metrics designed

to capture the amount of effort required to transform one distribution into another like Wasserstein, also known as the earth mover's distance. Others calculate the statistical significance of a distance between two distributions, such as Kullback-Leibler. However, they do not prove to be informative in the case of quantifying the difference between two flux distributions since they lead to p-values being too sensitive, resulting in extremely significant p-values for very similar distributions. In addition to this, these tests do provide scoring metrics, but they are unable to quantify or describe the differences with the goal of detecting flux shifts and change directions.

1.2 Challenges in assessing quality of predictions

Although SAMBA is a predictive method, evaluating the predictions using traditional contingency tables, recall and precision is difficult due to the nature of metabolomics measurements and the available "truth" datasets. The model contains all known metabolites involved in metabolic reactions, but metabolomics methods are not able to detect and annotate all of them. In fact, as it was shown in Frainay *et al.* [31], metabolites may be overlooked during the whole metabolomics pipeline. This can be for instance due to pre-processing steps since most peak picking methods [159] will define an intensity threshold to keep only intense peaks and, as a consequence, may discard peaks of interest that fall just below the threshold. This results in many cases where metabolites are predicted to be of interest while they are not detected by typical assays. In these cases, the predictions could be correct while being considered as a "false positive". Instead of using "false positive" to represent these predictions, we simply present the entire ranked prediction results in order to orient the user towards certain metabolites or metabolite classes. We then evaluate the method using true positive ranks and the list of the top most changed metabolites, some of which could be considered false positives, but could also be unmeasured

metabolites. An additional method of evaluating the statistical validity of the results is by running a hypergeometric test, to test the significance of obtaining the number of correct predictions, for different rank cut-offs. This is shown in Figure 30 and highlights that the number of experimental metabolites predicted in the top ranks is significant.

Conversely, metabolites in the original experimental results but not predicted as highly ranked by SAMBA could be due to inconsistencies in the model, whether they are due to errors or unknowns, or an incorrectly simulated metabolic condition. Additionally, extra care should be taken when analysing low-ranking metabolites as their z-scores are very similar to each other. This means that their specific order does not indicate much information about the extent of how they were affected by the perturbation, only that they were affected very little.

SAMBA is based on ranking z-score absolute values, meaning that the metabolites whose exchange fluxes (and by extension concentrations) are more likely to change will be considered first. There are of course metabolites whose concentrations can change very little and have extreme consequences on the rest of the metabolism, such as via enzyme regulation, or if they are limiting substrates for example.

Finally, this ranking system bypasses the issues that come with using flux values directly, and especially helps in choosing which metabolites to focus on first. The comparison of metabolic profile recommendations between different scenario simulations can be achieved by considering the top most changed metabolites and their ranks, as opposed to the raw flux values.

1.3 Current limits in metabolic modelling

Metabolic genes linked to reactions through GPR relationships are a simplification of the intricate system of gene expression. Reaction and enzyme

activities are simulated as a binary system, either on or off, depending on the validity of the GPR. In order to simulate a gene knockout, the gene rule needs to be invalidated, after which the reaction will not be active. There are of course more complex enzymatic variations which involve a reduction in enzyme activity or even an overexpression of an enzyme, which are more difficult to simulate. Instead of going through the GPR relationship to "turn off" an enzyme, the reaction's bounds can be directly modified to reduce or increase the maximum flux for example. The difficulty comes from the fact that a certain "reduction" value must be chosen, as well as deciding how to define this reduction, since an interval can be changed in multiple ways. For example, a 50% reduction could imply that the maximum reaction bound is divided by two, but it could also mean that both of the bound values are reduced by 25% towards the halfway point.

A major point to take into account when considering the use of metabolomics data is whether the metabolites can easily be mapped to databases using unique identifiers. This manual curation step is not absent even when using simulations to generate metabolic profiles: most GSMNs have some identifiers for genes, reactions and metabolites, but they are not always homogeneous across all entities, and can sometimes contain errors. When mapping genetic perturbations to the network, manual checks must be done to ensure that the final enzymatic disruption corresponds to the actual effect of the genetic mutation, due to the GPRs linking genes and reactions. Furthermore, once metabolic changes have been simulated, extracting information by mapping to external biological knowledge can be difficult due to inconsistent metabolite names and identifiers, once again often requiring a manual step to check the correspondence. This issue also renders validation and integration with experimental data slow and non-automatic. More broadly, this corresponds to the interoperability challenge of the FAIR (Findable, Accessible, Interoperable and Re-usable) policy in open science [160]. There is hence a need both for modelling and metabolomics

scientists to define strategies to facilitate the integration of metabolites in GSMN. This work could rely on ontologies to provide the necessary flexibility in metabolite and lipid mapping as proposed in [161].

Furthermore, there are of course many other non-metabolic genes which could still play a role in metabolism and metabolic regulation. These other genes are not always included in metabolic networks since their mechanisms are not always known, or they may not be seen as relevant to metabolism. Additionally, metabolites are known to regulate enzymes via different mechanisms such as competitive inhibition, allosteric inhibition, and allosteric activation. These feedback loops depend on the presence and concentration of different metabolites in cellular compartments, which cannot be estimated using traditional CBM methods due to the steady-state assumption. There have been efforts towards integrating regulatory networks with GSMNs, often limited to bacteria due to the smaller scale [162, 163, 164].

The measurement of pure standards of metabolites is essential to obtain the highest level of confidence in metabolite identification (level 1 according to Metabolomics Standard Initiative [33]). Selecting which standards to measure is by itself a challenge, since samples can contain thousands of metabolites. Hence, SAMBA can be used by laboratories to select which standards to acquire in the context of the disease under study. More broadly, the top ranked list can also be used to identify families of metabolites to study as a whole, such as by extending the panel of measurable metabolites during a metabolomics experiment.

The goal of this thesis was to simulate whole-body metabolic markers using a generic genome-scale model. From a physiological point of view these models may seem to be somewhat over simplified in that a single metabolic system is represented. However the examples used in this study are genetic diseases, therefore they affect the genome of all of the cells in the body. While gene expression can depend on organs and tissue regions, the hypothesis here is that experimentally observed metabolic profiles are a combination of metabolite

exports from all tissues connected to biofluids, which is why they can be equated to metabolic profiles predicted using a genome-scale network. However, the modulation of a tissue-specific biomarker may be predicted incorrectly if it is normally (biologically) compensated by other tissues, which could result in false positives. In those cases, tissue-specific networks could be useful for analysing diseases that are known to affect a certain tissue, such as the liver with glycogen storage diseases. These diseases are a collection of genetic metabolic disorders, and the enzymes affected by the mutations are specific to the liver and muscle [165]. By using transcriptomics data to create a liver-specific model, the accuracy of metabolic simulations could be increased. This can be done using various integration methods such as iMAT [166] or DEXOM [107]. However, choosing any given model and tissue-specific conditions must be done with care as it will have a major impact on the resulting metabolite ranks. More broadly, the definition of constraints is key to adapting the model to the biological condition (e.g. availability of nutrients) and will impact predictions. These modelling steps can be performed upstream of SAMBA.

The sheer quantity of results leads to a growing need for large-scale analysis methods of metabolomics results, and this is no different for simulated metabolic profiles. The analysis of pathway enrichment methods, whether applied to experimental or simulated data, is essential to knowing if the correct information is being extracted from this data.

2 Conclusion

Building upon constraint based modelling of metabolism through the use of random sampling of fluxes, we were able to predict large potential metabolic profiles and confirm measured metabolites both in targeted and untargeted assays. Ranking all metabolites becomes possible through the methodology's comparison of flux distributions between healthy and disease states. Metabolites

revealed by this method are of potential interest to broaden the panel of targets for future metabolomics experiments, and can be identified as understudied metabolites, helping to develop our understanding of metabolic mechanisms. Furthermore, the rank of a given metabolite can be compared between two different disruption scenarios, which provides information on the specificity of the disrupted metabolite to the scenario.

Although the methodology is designed to be used to predict external metabolite exchange fluxes, it can also be used to simulate the internal reaction fluxes, which can be useful for understanding internal metabolism along with external metabolites. Finally, simulated metabolic profiles can also be used to benchmark various analyses specific to metabolomics, such as pathway analysis, or other analyses which require lots of data like machine learning.

3 Perspectives

3.1 Aiding in a metabolomics workflow

Ideally, SAMBA will be used in conjunction with real metabolomics workflows and will be useful at multiple levels. First, many SAMBA simulations could be run for many different conditions pseudo-randomly, which could help identify potential avenues of interest if no particular area is favoured initially. Next, predicted metabolic profiles could help in choosing the sample for an experiment. Several tissue-specific models could be created and subjected to the same metabolic perturbation. SAMBA could then predict metabolic profiles for each model which could then be compared to an experimental plasma sample in order to determine the tissue which matches the most. This tissue could be chosen as the optimal sample type for future experiments. When designing an experiment, if the metabolic perturbation is known or somewhat described, SAMBA can predict potential metabolites to target, or a class of metabolites to

optimise the metabolomics setup for before even doing any experiments, saving time, reducing costs, and providing a less biased approach to both targeted and untargeted setups. This could provide suspected metabolites of interest for a given condition with more "weight", leading to experiments which may not have taken place otherwise. Metabolite identification could also benefit from SAMBA predictions: if extra metabolites are expected to be in the sample or to be significantly abundant, they could be identified where they previously would be overlooked. SAMBA predictions in parallel to an experimental setup can also help extend the global metabolic profile, improving enrichment methods and interpretation.

3.2 Predicting the toxicological effect of nitrous oxide

A possible application of SAMBA is the prediction of metabolite deregulations in response to toxicological effects. Toxicology is the study of the adverse effects of external substances on biological organisms, as well as the diagnosis and treatment of their exposure. As a concrete example, one could investigate nitrous oxide (laughing gas, N_2O), which is a chemical compound with medical uses in anaesthesia and pain reduction, as well as uses in fuel propellant and whipped cream. Since the 18th century, it has been used as a recreational drug, inducing a euphoric state, hallucinations and relaxation when inhaled [167]. The toxicity of nitrous oxide has been demonstrated to decrease methionine synthase's (MS) enzymatic activity in rats [168]. In humans, use of this drug leads to similar symptoms to those seen in vitamin B12 deficiency, resulting in a modified form of vitamin B12 which is unable to bind with the MS enzyme, inactivating it. Following this, individuals who use nitrous oxide take vitamin B12 to combat the neurological effects of this drug. This masks the biomarker MMA (methylmalonic acid), whose plasmatic increase in concentration is traditionally used for the diagnosis of various vitamin B12 deficiencies. This

biomarker is also used to diagnose nitrous oxide intoxication, but is unable to be used as a biomarker when the patient has been automedicating with vitamin B12 supplements [169]. Figure 60 shows these known and possible effects in red.

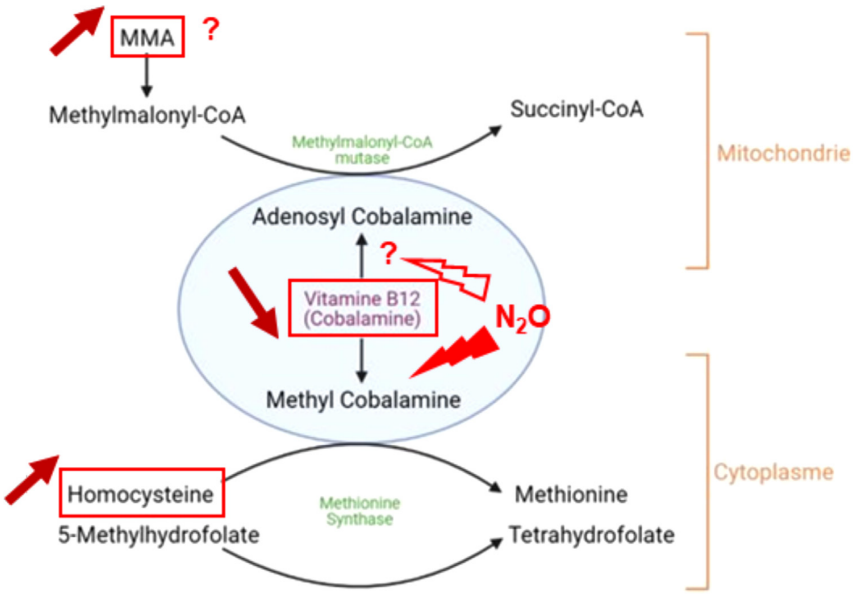


Figure 60: Known and possible impacts of nitrous oxide on its enzymatic targets and associated biological modifications. Figure from [169].

The reduction in MS enzyme activity could be simulated in a human model in order to reveal its effect in humans on multiple fronts (shown as red arrows and question marks in Figure 60).

- First, regarding the use of MMA as a biomarker for this condition: we could demonstrate that its increase is masked when the network is also supplied with vitamin B12.
- Second, we could demonstrate that MMA is not a biomarker specific to nitrous oxide intoxication, as suggested in [169], by simulating other conditions that lead to an increase in MMA.
- Third, plasmatic homocysteine has been shown to increase during nitrous oxide intoxication even when supplemented with vitamin B12 [170]. This could be simulated in order to demonstrate its potential use as a biomarker for suspected nitrous oxide intoxication.

- Fourth, MMA-CoA-mutase, a mitochondrial enzyme which uses MMA-CoA, is often said to be affected by the nitrous oxide-induced vitamin B12 modification, but this direct link has not been demonstrated, and it has even been shown to be unaffected in rats [171]. The fluxes of this enzyme could be simulated in humans in the previously described model.

3.3 Beyond single knockout scenarios

Furthermore, including the SAMBA approach in whole-body metabolic models [111] which combines the interactions of multiple human tissues is a potential path for future study. Since sampling algorithms are being continuously improved and iterated upon, and more CPU power is being added to computational clusters, running sampling on these larger models will become less of an issue. These models, with their different gene and reaction expressions per tissue, could reveal the different effects of genetic diseases or other metabolic disruptions on biofluid metabolites on a multi-tissular level.

In order to continue to highlight the full benefits of using sampling distributions instead of FVA boundary values, further research for other applications and more validation data are required. For instance, *in vivo* fluxomics data [172] measured experimentally could be matched to simulated import/export rates. Other sampling algorithms for GSMNs exist and could also improve predictions if applied correctly, and there is much more to explore when it comes to improving the scoring metric for sampling-based metabolic changes.

Instead of sampling the exchange reactions, internal fluxes could also be investigated in relation to both enrichment methods and metabolic profile predictions. This could provide insight into how other reactions and pathways are disrupted in any given specific metabolic condition and help describe the effects across the entire network.

Finally, while SAMBA was applied to KO scenarios in this paper, the method

can be adapted to more complex constraints such as multiple gene KOs or even to simulate knock downs of reactions. Knock downs involve reducing the maximum flux capacity of affected reactions instead of blocking the flux completely and can be run directly using SAMBA by changing the input condition file (see Figure 26 for details). This can be particularly useful in the context of toxicology or drug development, where these subtle metabolic disruptions can lead to reduced enzyme activity. There are many potential applications for SAMBA recommendations, such as in predicting the effects of xenobiotics on human metabolism. In this thesis, the focus was on simulating genetic diseases as the metabolic disruptions are simple to translate into the metabolic model, but the next challenge will be converting more complex metabolic perturbations into explicit reaction modulations. Effects like toxic environmental exposure can be simulated once the mechanism is narrowed down, while the effect of diet could be modelled by varying the input nutrients via the exchange reactions of the network.

3.4 Generating databases of simulated metabolic profiles

From a more long-term point of view, the next steps for the use of the SAMBA pipeline involve benefitting from the advantages of simulated data, by generating many scenarios and many corresponding metabolic profiles. These large quantities of associated known metabolic perturbations and lists of metabolites can be used in multiple ways.

The first goal is to make use of knowing where the metabolic perturbation lies in order to benchmark pathway enrichment methods, continuing the collaborative work from Chapter IV. This will consist of generating more scenarios of metabolic disruptions and developing ways to push the analysis of pathway enrichment methods further, through statistical, analytical and graph-based techniques. Once the biases involved with pathway enrichment

methods for metabolomics data are fully known, the community can become more aware via precautions and extra guidelines, and perhaps more metabolomics-appropriate methods will be developed.

Furthermore, another approach to using SAMBA's generative power to its fullest is by creating a simulated database of every unique KO, KD, multiple KOs, and combinations of many different disruptions, and compiling it as a repository for comparison with real data. This could be used in conjunction with biological data to determine which metabolic perturbations are most likely to cause the condition tested by the experiment, by matching the experimental profile with the simulated metabolic profile.

An in depth example of how this could by applied to a biological scenario is the field of toxicology. As we become increasingly exposed to a multitude of chemical substances, it is essential to comprehend the potential toxicological consequences on human health and the environment. By entering the body, these molecules can disrupt gene expression, cell functions, metabolism and hormone regulation systems. A class of compounds called endocrine disrupting chemicals are known to affect both metabolism and interfere with hormones (endocrine system), and are suspected to cause increased risk of obesity [173], diabetes [174], diminished immune systems [175], cancer [176], with more potential understudied effects [177, 178].

Traditional toxicological studies rely on the assessment of individual chemicals and their effects on specific organs or systems. By applying metabolomics to toxicology, a new way of comprehensively investigating the perturbations induced by toxic compounds is available at the molecular level [179]. By monitoring changes in many metabolite levels at once, metabolomics presents an opportunity to understand intricate cellular interactions and the molecular mechanisms underlying toxicity responses [180]. The main issue is that these metabolic and molecular mechanisms are poorly described as they are difficult to reproduce in laboratory conditions and, for obvious ethical reasons,

no intervention studies can be performed on humans to assess the toxicological impacts of chemicals. This is often due to the long-term exposure required to see physiological effects on an individual, tissue or sample, as well as the complex interconnected processes involving multiple tissues or systems (hormonal and metabolic for example).

Using a simulated database could help pinpoint the modes of action of xenobiotics (compounds extrinsic to an organism) involving metabolism. By mapping experimental metabolic profiles obtained from an exposed sample to predicted profiles using random sampling in a metabolic network, the affected area(s) can be highlighted since the metabolic perturbation is known when simulating a metabolic profile. The most probable areas of metabolism to have caused the metabolic profile can therefore be identified, providing potential targets of interest for future experiments and analysis.

Bibliography

[1] Ada Hamosh *et al.* "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders". In: *Nucleic Acids Research* 33.Database Issue (Jan. 2005), pp. D514–D517. ISSN: 0305-1048. DOI: 10.1093/nar/gki033. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC539987/ (Cited on pages 25, 66, 107).

[2] A. M. DiGeorge *et al.* "Prospective study of maple-syrup-urine disease for the first four days of life". eng. In: *The New England Journal of Medicine* 307.24 (Dec. 1982), pp. 1492–1495. ISSN: 0028-4793. DOI: 10.1056/NEJM198212093072405 (Cited on page 26).

[3] Christopher J. Clarke and John N. Haselden. "Metabolic profiling as a tool for understanding mechanisms of toxicity". eng. In: *Toxicologic Pathology* 36.1 (Jan. 2008), pp. 140–147. ISSN: 1533-1601. DOI: 10.1177/0192623307310947 (Cited on page 26).

[4] Fabienne Jeanneret *et al.* "Evaluation and identification of dioxin exposure biomarkers in human urine by high-resolution metabolomics, multivariate analysis and in vitro synthesis". In: *Toxicology Letters* 240.1 (Jan. 2016), pp. 22–31. ISSN: 0378-4274. DOI: 10.1016/j.toxlet.2015.10.004. URL: https://www.sciencedirect.com/science/article/pii/S0378427415300680 (Cited on page 26).

[5] Horace R. T. Williams *et al.* "Serum Metabolic Profiling in Inflammatory Bowel Disease". en. In: *Digestive Diseases and Sciences* 57.8 (Aug. 2012),

pp. 2157–2165. ISSN: 1573-2568. DOI: 10.1007/s10620-012-2127-2. URL: https://doi.org/10.1007/s10620-012-2127-2 (Cited on page 27).

[6] Karl Burgess and Naomi Rankin. "Metabolomics for the diagnosis of influenza". In: *EBioMedicine* 72 (Oct. 2021), p. 103599. ISSN: 2352-3964. DOI: 10.1016/j.ebiom.2021.103599. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8501667/ (Cited on page 27).

[7] Cristina Menni *et al.* "Targeted metabolomics profiles are strongly correlated with nutritional patterns in women". en. In: *Metabolomics* 9.2 (Apr. 2013), pp. 506–514. ISSN: 1573-3890. DOI: 10.1007/s11306-012-0469-6. URL: https://doi.org/10.1007/s11306-012-0469-6 (Cited on page 27).

[8] Audrey LeGouëllec *et al.* "High-Resolution Magic Angle Spinning NMR-Based Metabolomics Revealing Metabolic Changes in Lung of Mice Infected with P. aeruginosa Consistent with the Degree of Disease Severity." eng. In: *Journal of proteome research* 17.10 (Oct. 2018). Place: United States, pp. 3409–3417. ISSN: 1535-3907 1535-3893. DOI: 10.1021/acs.jproteome.8b00306 (Cited on page 27).

[9] Oliver Fiehn *et al.* "Metabolite profiling for plant functional genomics". en. In: *Nature Biotechnology* 18.11 (Nov. 2000). Number: 11 Publisher: Nature Publishing Group, pp. 1157–1161. ISSN: 1546-1696. DOI: 10.1038/81137. URL: https://www.nature.com/articles/nbt1100_1157 (Cited on page 27).

[10] Yang Li *et al.* "A Functional Genomics Approach to Understand Variation in Cytokine Production in Humans". In: *Cell* 167.4 (Nov. 2016), 1099–1110.e14. ISSN: 0092-8674. DOI: 10.1016/j.cell.2016.10.017. URL: https://www.sciencedirect.com/science/article/pii/S0092867416314003 (Cited on page 27).

[11] Alan F Wright. "Genetic Variation: Polymorphisms and Mutations". en. In: *Encyclopedia of Life Sciences*. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1038/npg John Wiley & Sons, Ltd, 2005. ISBN: 978-0-470-01590-2. DOI: 10.1038/npg.els.0005005. URL: https://onlinelibrary.wiley.com/doi/abs/10.1038/npg.els.0005005 (Cited on page 27).

[12] Chia-Chen Liu *et al.* "Apolipoprotein E and Alzheimer disease: risk, mechanisms, and therapy". In: *Nature reviews. Neurology* 9.2 (Feb. 2013), pp. 106–118. ISSN: 1759-4758. DOI: 10.1038/nrneurol.2012.263. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3726719/ (Cited on page 27).

[13] Gilbert S. Omenn. "Reflections on the HUPO Human Proteome Project, the Flagship Project of the Human Proteome Organization, at 10 Years". In: *Molecular & Cellular Proteomics* 20 (Jan. 2021), p. 100062. ISSN: 1535-9476. DOI: 10.1016/j.mcpro.2021.100062. URL: https://www.sciencedirect.com/science/article/pii/S1535947621000359 (Cited on page 28).

[14] Emil Uffelmann *et al.* "Genome-wide association studies". en. In: *Nature Reviews Methods Primers* 1.1 (Aug. 2021). Number: 1 Publisher: Nature Publishing Group, pp. 1–21. ISSN: 2662-8449. DOI: 10.1038/s43586-021-00056-9. URL: https://www.nature.com/articles/s43586-021-00056-9 (Cited on page 28).

[15] Christian Gieger *et al.* "Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum". eng. In: *PLoS genetics* 4.11 (Nov. 2008), e1000282. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1000282 (Cited on page 28).

[16] Shuangfeng Yang *et al.* "mGWAS identification of six novel single nucleotide polymorphism loci with strong correlation to gastric cancer". In: *Cancer & Metabolism* 9.1 (Sept. 2021), p. 34. ISSN: 2049-3002. DOI:

10.1186/s40170-021-00269-2. URL: https://doi.org/10.1186/s40170-021-00269-2 (Cited on page 28).

[17] Daniel Montemayor and Kumar Sharma. "mGWAS: next generation genetic prediction in kidney disease". In: *Nature reviews. Nephrology* 16.5 (May 2020), pp. 255–256. ISSN: 1759-5061. DOI: 10.1038/s41581-020-0270-0. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7261371/ (Cited on page 28).

[18] Marianne L. Slaten *et al.* "mGWAS Uncovers Gln-Glucosinolate Seed-Specific Interaction and its Role in Metabolic Homeostasis1 [OPEN]". In: *Plant Physiology* 183.2 (June 2020), pp. 483–500. ISSN: 0032-0889. DOI: 10.1104/pp.20.00039. URL: https://doi.org/10.1104/pp.20.00039 (Cited on page 29).

[19] Priscilla L. Yang. "Metabolomics and Lipidomics". In: *Viral Pathogenesis* (2016), pp. 181–198. DOI: 10.1016/B978-0-12-800964-2.00014-8. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7149616/ (Cited on page 30).

[20] Abdul-Hamid Emwas *et al.* "Fluxomics - New Metabolomics Approaches to Monitor Metabolic Pathways". In: *Frontiers in Pharmacology* 13 (2022). ISSN: 1663-9812. URL: https://www.frontiersin.org/articles/10.3389/fphar.2022.805782 (Cited on page 31).

[21] Brian P. Lankadurai, Edward G. Nagato, and Myrna J. Simpson. "Environmental metabolomics: an emerging approach to study organism responses to environmental stressors". In: *Environmental Reviews* 21.3 (Sept. 2013). Publisher: NRC Research Press, pp. 180–205. ISSN: 1181-8700. DOI: 10.1139/er-2013-0011. URL: https://cdnsciencepub.com/doi/full/10.1139/er-2013-0011 (Cited on page 32).

[22] William J. Griffiths *et al.* "Targeted metabolomics for biomarker discovery". eng. In: *Angewandte Chemie (International Ed. in English)* 49.32 (July 2010), pp. 5426–5445. ISSN: 1521-3773. DOI: 10.1002/anie.200905579 (Cited on page 32).

[23] Daisuke Saigusa *et al.* "Identification of biomarkers to diagnose diseases and find adverse drug reactions by metabolomics". In: *Drug Metabolism and Pharmacokinetics* 37 (Apr. 2021), p. 100373. ISSN: 1347-4367. DOI: 10.1016/j.dmpk.2020.11.008. URL: https://www.sciencedirect.com/science/article/pii/S1347436720304341 (Cited on page 32).

[24] Judith JM. Jans, Melissa H. Broeks, and Nanda M. Verhoeven-Duif. "Metabolomics in diagnostics of inborn metabolic disorders". In: *Current Opinion in Systems Biology* 29 (Mar. 2022), p. 100409. ISSN: 2452-3100. DOI: 10.1016/j.coisb.2021.100409. URL: https://www.sciencedirect.com/science/article/pii/S2452310021001049 (Cited on page 32).

[25] Vivian Tounta *et al.* "Metabolomics in infectious diseases and drug discovery". In: *Molecular Omics* 17.3 (), pp. 376–393. ISSN: 2515-4184. DOI: 10.1039/d1mo00017a. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8202295/ (Cited on page 32).

[26] David S. Wishart. "Emerging applications of metabolomics in drug discovery and precision medicine". en. In: *Nature Reviews Drug Discovery* 15.7 (July 2016). Number: 7 Publisher: Nature Publishing Group, pp. 473–484. ISSN: 1474-1784. DOI: 10.1038/nrd.2016.32. URL: https://www.nature.com/articles/nrd.2016.32 (Cited on page 32).

[27] David S. Wishart. "Metabolomics: applications to food science and nutrition research". In: *Trends in Food Science & Technology* 19.9 (Sept. 2008), pp. 482–493. ISSN: 0924-2244. DOI: 10.1016/j.tifs.2008.03.003. URL: https://www.sciencedirect.com/science/article/pii/S0924224408000770 (Cited on page 32).

[28] Jun Feng Xiao, Bin Zhou, and Habtom W. Ressom. "Metabolite identification and quantitation in LC-MS/MS-based metabolomics". In: *Trends in analytical chemistry : TRAC* 32 (Feb. 2012), pp. 1–14. ISSN: 0165-9936. DOI: 10.1016/j.trac.2011.08.009. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3278153/ (Cited on page 33).

[29] Georgios Theodoridis, Helen G. Gika, and Ian D. Wilson. "LC-MS-based methodology for global metabolite profiling in metabonomics/metabolomics". en. In: *TrAC Trends in Analytical Chemistry*. Metabolomics 27.3 (Mar. 2008), pp. 251–260. ISSN: 0165-9936. DOI: 10.1016/j.trac.2008.01.008. URL: https://www.sciencedirect.com/science/article/pii/S0165993608000095 (Cited on page 33).

[30] Mingxun Wang *et al.* "Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking". en. In: *Nature Biotechnology* 34.8 (Aug. 2016). Number: 8 Publisher: Nature Publishing Group, pp. 828–837. ISSN: 1546-1696. DOI: 10.1038/nbt.3597. URL: https://www.nature.com/articles/nbt.3597 (Cited on page 34).

[31] Clément Frainay *et al.* "Mind the Gap: Mapping Mass Spectral Databases in Genome-Scale Metabolic Networks Reveals Poorly Covered Areas". eng. In: *Metabolites* 8.3 (Sept. 2018), E51. ISSN: 2218-1989. DOI: 10.3390/metabo8030051 (Cited on pages 35, 169).

[32] Svati H. Shah, William E. Kraus, and Christopher B. Newgard. "Metabolomic Profiling for Identification of Novel Biomarkers and Mechanisms Related to Common Cardiovascular Diseases: Form and Function". In: *Circulation* 126.9 (Aug. 2012), pp. 1110–1120. ISSN: 0009-7322. DOI: 10.1161/CIRCULATIONAHA.111.060368. URL: https:

//www.ncbi.nlm.nih.gov/pmc/articles/PMC4374548/ (Cited on pages 35, 38).

[33] Lloyd W. Sumner *et al.* "Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI)". In: *Metabolomics : Official journal of the Metabolomic Society* 3.3 (Sept. 2007), pp. 211–221. ISSN: 1573-3882. DOI: 10.1007/s11306-007-0082-2. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3772505/ (Cited on pages 36, 172).

[34] Emma L. Schymanski *et al.* "Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence". In: *Environmental Science & Technology* 48.4 (Feb. 2014). Publisher: American Chemical Society, pp. 2097–2098. ISSN: 0013-936X. DOI: 10.1021/es5002105. URL: https://doi.org/10.1021/es5002105 (Cited on page 36).

[35] Minoru Kanehisa *et al.* "KEGG for taxonomy-based analysis of pathways and genomes". eng. In: *Nucleic Acids Research* 51.D1 (Jan. 2023), pp. D587–D592. ISSN: 1362-4962. DOI: 10.1093/nar/gkac963 (Cited on page 42).

[36] M. Kanehisa and S. Goto. "KEGG: kyoto encyclopedia of genes and genomes". eng. In: *Nucleic Acids Research* 28.1 (Jan. 2000), pp. 27–30. ISSN: 0305-1048. DOI: 10.1093/nar/28.1.27 (Cited on page 42).

[37] Minoru Kanehisa. "Toward understanding the origin and evolution of cellular organisms". eng. In: *Protein Science: A Publication of the Protein Society* 28.11 (Nov. 2019), pp. 1947–1951. ISSN: 1469-896X. DOI: 10.1002/pro.3715 (Cited on page 42).

[38] Ron Caspi, Kate Dreher, and Peter D. Karp. "The challenge of constructing, classifying and representing metabolic pathways". In: *FEMS microbiology letters* 345.2 (Aug. 2013), pp. 85–93. ISSN: 0378-1097.

DOI: 10.1111/1574-6968.12194. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4026850/ (Cited on page 42).

[39] Ron Caspi *et al.* "The MetaCyc database of metabolic pathways and enzymes - a 2019 update". eng. In: *Nucleic Acids Research* 48.D1 (Jan. 2020), pp. D445–D453. ISSN: 1362-4962. DOI: 10.1093/nar/gkz862 (Cited on page 42).

[40] Peter D. Karp and Ron Caspi. "A Survey of Metabolic Databases Emphasizing the MetaCyc Family". In: *Archives of Toxicology* 85.9 (Sept. 2011), pp. 1015–1033. ISSN: 0340-5761. DOI: 10.1007/s00204-011-0705-2. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3352032/ (Cited on page 42).

[41] Marc Gillespie *et al.* "The reactome pathway knowledgebase 2022". In: *Nucleic Acids Research* 50.D1 (Jan. 2022), pp. D687–D692. ISSN: 0305-1048. DOI: 10.1093/nar/gkab1028. URL: https://doi.org/10.1093/nar/gkab1028 (Cited on pages 42, 48).

[42] Parit Bansal *et al.* "Rhea, the reaction knowledgebase in 2022". In: *Nucleic Acids Research* 50.D1 (Jan. 2022), pp. D693–D700. ISSN: 0305-1048. DOI: 10.1093/nar/gkab1016. URL: https://doi.org/10.1093/nar/gkab1016 (Cited on page 43).

[43] Cecilia Wieder *et al.* "Pathway analysis in metabolomics: Recommendations for the use of over-representation analysis". In: *PLoS Computational Biology* 17.9 (Sept. 2021), e1009105. ISSN: 1553-734X. DOI: 10.1371/journal.pcbi.1009105. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8448349/ (Cited on pages 43, 145, 148–150).

[44] Hagai Ginsburg. "Caveat emptor: limitations of the automated reconstruction of metabolic pathways in Plasmodium". In: *Trends in Parasitology* 25.1 (Jan. 2009), pp. 37–43. ISSN: 1471-4922. DOI:

10.1016/j.pt.2008.08.012. URL: https://www.sciencedirect.com/science/article/pii/S1471492208002535 (Cited on pages 43, 147).

[45] Fergal J Martin *et al.* "Ensembl 2023". In: *Nucleic Acids Research* 51.D1 (Jan. 2023), pp. D933–D941. ISSN: 0305-1048. DOI: 10.1093/nar/gkac958. URL: https://doi.org/10.1093/nar/gkac958 (Cited on page 43).

[46] Eric W. Sayers *et al.* "Database resources of the national center for biotechnology information". eng. In: *Nucleic Acids Research* 50.D1 (Jan. 2022), pp. D20–D26. ISSN: 1362-4962. DOI: 10.1093/nar/gkab1112 (Cited on page 43).

[47] International Union of Biochemistry and Molecular Biology Nomenclature Committee and Edwin C. Webb. *Enzyme Nomenclature 1992: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology and the Nomenclature and Classification of Enzymes*. en. Google-Books-ID: T9BSd7PyS6AC. Elsevier Science, Aug. 1992. ISBN: 978-0-12-227165-6 (Cited on page 43).

[48] The UniProt Consortium. "UniProt: the Universal Protein Knowledgebase in 2023". In: *Nucleic Acids Research* 51.D1 (Jan. 2023), pp. D523–D531. ISSN: 0305-1048. DOI: 10.1093/nar/gkac1052. URL: https://doi.org/10.1093/nar/gkac1052 (Cited on page 43).

[49] Janna Hastings *et al.* "ChEBI in 2016: Improved services and an expanding collection of metabolites". eng. In: *Nucleic acids research* 44.D1 (Jan. 2016), pp. D1214–9. ISSN: 1362-4962. DOI: 10.1093/nar/gkv1031. URL: https://europepmc.org/articles/PMC4702775 (Cited on pages 43, 50).

[50] Sunghwan Kim *et al.* "PubChem 2023 update". In: *Nucleic Acids Research* 51.D1 (Jan. 2023), pp. D1373–D1380. ISSN: 0305-1048. DOI: 10.1093/nar/gkac956. URL: https://doi.org/10.1093/nar/gkac956 (Cited on page 43).

[51] David S. Wishart *et al.* "HMDB: the Human Metabolome Database". eng. In: *Nucleic Acids Research* 35.Database issue (Jan. 2007), pp. D521–526. ISSN: 1362-4962. DOI: 10.1093/nar/gkl923 (Cited on page 43).

[52] David Weininger. "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules". In: *Journal of Chemical Information and Computer Sciences* 28.1 (Feb. 1988). Publisher: American Chemical Society, pp. 31–36. ISSN: 0095-2338. DOI: 10.1021/ci00057a005. URL: https://doi.org/10.1021/ci00057a005 (Cited on page 43).

[53] Stephen R. Heller *et al.* "InChI, the IUPAC International Chemical Identifier". In: *Journal of Cheminformatics* 7.1 (May 2015), p. 23. ISSN: 1758-2946. DOI: 10.1186/s13321-015-0068-4. URL: https://doi.org/10.1186/s13321-015-0068-4 (Cited on page 43).

[54] Ville Koistinen *et al.* "Towards a Rosetta stone for metabolomics: recommendations to overcome inconsistent metabolite nomenclature". en. In: *Nature Metabolism* 5.3 (Mar. 2023). Number: 3 Publisher: Nature Publishing Group, pp. 351–354. ISSN: 2522-5812. DOI: 10.1038/s42255-023-00757-3. URL: https://www.nature.com/articles/s42255-023-00757-3 (Cited on page 43).

[55] Zhiqiang Pang *et al.* "MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights". In: *Nucleic Acids Research* 49.W1 (July 2021), W388–W396. ISSN: 0305-1048. DOI: 10.1093/nar/gkab382. URL: https://doi.org/10.1093/nar/gkab382 (Cited on page 44).

[56] Zachary A. King *et al.* "BiGG Models: A platform for integrating, standardizing and sharing genome-scale models". In: *Nucleic Acids Research* 44.D1 (Jan. 2016), pp. D515–D522. ISSN: 0305-1048. DOI: 10.1093/nar/gkv1049. URL: https://doi.org/10.1093/nar/gkv1049 (Cited on page 44).

[57] Sébastien Moretti *et al.* "MetaNetX/MNXref: unified namespace for metabolites and biochemical reactions in the context of metabolic models". In: *Nucleic Acids Research* 49.D1 (Jan. 2021), pp. D570–D574. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa992. URL: https://doi.org/10.1093/nar/gkaa992 (Cited on page 44).

[58] Rahuman S Malik-Sheriff *et al.* "BioModels—15 years of sharing computational models in life science". In: *Nucleic Acids Research* 48.D1 (Jan. 2020), pp. D407–D415. ISSN: 0305-1048. DOI: 10.1093/nar/gkz1055. URL: https://doi.org/10.1093/nar/gkz1055 (Cited on page 44).

[59] Feiran Li *et al.* "GotEnzymes: an extensive database of enzyme parameter predictions". In: *Nucleic Acids Research* 51.D1 (Jan. 2023), pp. D583–D586. ISSN: 0305-1048. DOI: 10.1093/nar/gkac831. URL: https://doi.org/10.1093/nar/gkac831 (Cited on page 44).

[60] Jonathan L. Robinson *et al.* "An atlas of human metabolism". eng. In: *Science Signaling* 13.624 (Mar. 2020), eaaz1482. ISSN: 1937-9145. DOI: 10.1126/scisignal.aaz1482 (Cited on pages 44, 49, 110).

[61] Hao Wang *et al.* "Genome-scale metabolic network reconstruction of model animals as a platform for translational research". In: *Proceedings of the National Academy of Sciences* 118.30 (July 2021). Publisher: Proceedings of the National Academy of Sciences, e2102344118. DOI: 10.1073/pnas.2102344118. URL: https://www.pnas.org/doi/full/10.1073/pnas.2102344118 (Cited on page 44).

[62] M. Hucka *et al.* "The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models". In: *Bioinformatics* 19.4 (Mar. 2003), pp. 524–531. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btg015. URL: https://doi.org/10.1093/bioinformatics/btg015 (Cited on page 45).

[63] A. Finney and M. Hucka. "Systems biology markup language: Level 2 and beyond". eng. In: *Biochemical Society Transactions* 31.Pt 6 (Dec. 2003), pp. 1472–1473. ISSN: 0300-5127. DOI: 10.1042/bst0311472 (Cited on page 45).

[64] Michael Hucka *et al.* "The Systems Biology Markup Language (SBML): Language Specification for Level 3 Version 1 Core". In: *Journal of integrative bioinformatics* 12.2 (Sept. 2015), p. 266. ISSN: 1613-4516. DOI: 10.2390/biecoll-jib-2015-266. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5451324/ (Cited on page 45).

[65] Oliver Hädicke and Steffen Klamt. "EColiCore2: a reference network model of the central metabolism of Escherichia coli and relationships to its genome-scale parent model". eng. In: *Scientific Reports* 7 (Jan. 2017), p. 39647. ISSN: 2045-2322. DOI: 10.1038/srep39647 (Cited on page 46).

[66] Martin I. Sigurdsson *et al.* "A detailed genome-wide reconstruction of mouse metabolism based on human Recon 1". eng. In: *BMC systems biology* 4 (Oct. 2010), p. 140. ISSN: 1752-0509. DOI: 10.1186/1752-0509-4-140 (Cited on page 46).

[67] Ines Thiele and Bernhard Ø Palsson. "A protocol for generating a high-quality genome-scale metabolic reconstruction". en. In: *Nature Protocols* 5.1 (Jan. 2010). Number: 1 Publisher: Nature Publishing Group, pp. 93–121. ISSN: 1750-2799. DOI: 10.1038/nprot.2009.203. URL: https://www.nature.com/articles/nprot.2009.203 (Cited on page 46).

[68] Arnaud Belcour. "Combining knowledge-based and sequence comparison approaches to elucidate metabolic functions, from pathways to communities". These de doctorat. Rennes 1, Oct. 2022. URL: https://www.theses.fr/2022REN1S061 (Cited on page 47).

[69] Jinliang Chen *et al.* "Modelling predicts tomatoes can be bigger and sweeter if biophysical factors and transmembrane transports are fine-tuned during fruit development". en. In: *New Phytologist* 230 (2021), p. 1489. DOI: 10.1111/nph.17260. URL: https://hal.inrae.fr/hal-03152042 (Cited on page 48).

[70] Tungadri Bose *et al.* "Understanding the role of interactions between host and Mycobacterium tuberculosis under hypoxic condition: an in silico approach". eng. In: *BMC genomics* 19.1 (July 2018), p. 555. ISSN: 1471-2164. DOI: 10.1186/s12864-018-4947-8 (Cited on page 48).

[71] Gabriela I. Guzmán *et al.* "Model-driven discovery of underground metabolic functions in Escherichia coli". In: *Proceedings of the National Academy of Sciences* 112.3 (Jan. 2015). Publisher: Proceedings of the National Academy of Sciences, pp. 929–934. DOI: 10.1073/pnas.1414218112. URL: https://www.pnas.org/doi/full/10.1073/pnas.1414218112 (Cited on page 48).

[72] Alan R. Pacheco, Mauricio Moel, and Daniel Segrè. "Costless metabolic secretions as drivers of interspecies interactions in microbial ecosystems". en. In: *Nature Communications* 10.1 (Jan. 2019). Number: 1 Publisher: Nature Publishing Group, p. 103. ISSN: 2041-1723. DOI: 10.1038/s41467-018-07946-9. URL: https://www.nature.com/articles/s41467-018-07946-9 (Cited on page 48).

[73] Wonhee Hur *et al.* "Systems approach to characterize the metabolism of liver cancer stem cells expressing CD133". en. In: *Scientific Reports* 7.1 (Apr. 2017). Number: 1 Publisher: Nature Publishing Group, p. 45557. ISSN: 2045-2322. DOI: 10.1038/srep45557. URL: https://www.nature.com/articles/srep45557 (Cited on page 48).

[74] Yazdan Asgari *et al.* "Exploring candidate biomarkers for lung and prostate cancers using gene expression and flux variability analysis".

In: *Integrative Biology* 10.2 (Feb. 2018), pp. 113–120. ISSN: 1757-9708. DOI: 10.1039/c7ib00135e. URL: https://doi.org/10.1039/c7ib00135e (Cited on page 48).

[75] Robert D. Fleischmann *et al.* "Whole-Genome Random Sequencing and Assembly of Haemophilus influenzae Rd". In: *Science* 269.5223 (July 1995). Publisher: American Association for the Advancement of Science, pp. 496–512. DOI: 10.1126/science.7542800. URL: https://www.science.org/doi/10.1126/science.7542800 (Cited on page 48).

[76] J. S. Edwards and B. O. Palsson. "The Escherichia coli MG1655 in silico metabolic genotype: Its definition, characteristics, and capabilities". In: *Proceedings of the National Academy of Sciences of the United States of America* 97.10 (May 2000), pp. 5528–5533. ISSN: 0027-8424. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC25862/ (Cited on page 48).

[77] THE C. ELEGANS SEQUENCING CONSORTIUM. "Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology". In: *Science* 282.5396 (Dec. 1998). Publisher: American Association for the Advancement of Science, pp. 2012–2018. DOI: 10.1126/science.282.5396.2012. URL: https://www.science.org/doi/10.1126/science.282.5396.2012 (Cited on page 48).

[78] International Human Genome Sequencing Consortium. "Finishing the euchromatic sequence of the human genome". eng. In: *Nature* 431.7011 (Oct. 2004), pp. 931–945. ISSN: 1476-4687. DOI: 10.1038/nature03001 (Cited on page 48).

[79] Pedro Romero *et al.* "Computational prediction of human metabolic pathways from the complete human genome". In: *Genome Biology* 6.1 (Dec. 2004), R2. ISSN: 1474-760X. DOI: 10.1186/gb-2004-6-1-r2. URL: https://doi.org/10.1186/gb-2004-6-1-r2 (Cited on page 48).

[80] G. Joshi-Tope *et al.* "Reactome: a knowledgebase of biological pathways". In: *Nucleic Acids Research* 33.Database Issue (Jan. 2005), pp. D428–D432. ISSN: 0305-1048. DOI: 10.1093/nar/gki072. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC540026/ (Cited on page 48).

[81] Natalie C. Duarte *et al.* "Global reconstruction of the human metabolic network based on genomic and bibliomic data". In: *Proceedings of the National Academy of Sciences of the United States of America* 104.6 (Feb. 2007), pp. 1777–1782. ISSN: 0027-8424. DOI: 10.1073/pnas.0610772104. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1794290/ (Cited on pages 48, 66).

[82] Hongwu Ma *et al.* "The Edinburgh human metabolic network reconstruction and its functional analysis". In: *Molecular Systems Biology* 3 (Sept. 2007), p. 135. ISSN: 1744-4292. DOI: 10.1038/msb4100177. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2013923/ (Cited on page 48).

[83] Ines Thiele *et al.* "A community-driven global reconstruction of human metabolism". en. In: *Nature Biotechnology* 31.5 (May 2013). Number: 5 Publisher: Nature Publishing Group, pp. 419–425. ISSN: 1546-1696. DOI: 10.1038/nbt.2488. URL: https://www.nature.com/articles/nbt.2488 (Cited on pages 48, 66, 98, 107, 108, 110, 168).

[84] Christoph Gille *et al.* "HepatoNet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology". In: *Molecular Systems Biology* 6 (Sept. 2010), p. 411. ISSN: 1744-4292. DOI: 10.1038/msb.2010.62. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2964118/ (Cited on page 48).

[85] Neil Swainston *et al.* "Recon 2.2: from reconstruction to model of human metabolism". In: *Metabolomics* 12 (2016), p. 109. ISSN: 1573-3882. DOI: 10.

1007/s11306-016-1051-4. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4896983/ (Cited on page 49).

[86] Adil Mardinoglu *et al.* "Integration of clinical data with a genome-scale metabolic model of the human adipocyte". In: *Molecular Systems Biology* 9 (Mar. 2013), p. 649. ISSN: 1744-4292. DOI: 10.1038/msb.2013.5. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3619940/ (Cited on page 49).

[87] Natapol Pornputtapong, Intawat Nookaew, and Jens Nielsen. "Human metabolic atlas: an online resource for human metabolism". In: *Database: The Journal of Biological Databases and Curation* 2015 (July 2015), bav068. ISSN: 1758-0463. DOI: 10.1093/database/bav068. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4513696/ (Cited on page 49).

[88] Adil Mardinoglu *et al.* "Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease". eng. In: *Nature Communications* 5 (2014), p. 3083. ISSN: 2041-1723. DOI: 10.1038/ncomms4083 (Cited on page 49).

[89] Edik M. Blais *et al.* "Reconciled rat and human metabolic networks for comparative toxicogenomics and biomarker predictions". In: *Nature Communications* 8 (Feb. 2017), p. 14250. ISSN: 2041-1723. DOI: 10.1038/ncomms14250. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5309818/ (Cited on page 49).

[90] Elizabeth Brunk *et al.* "Recon3D: A Resource Enabling A Three-Dimensional View of Gene Variation in Human Metabolism". en. In: *Nature biotechnology* 36.3 (Mar. 2018). Publisher: NIH Public Access, p. 272. DOI: 10.1038/nbt.4072. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5840010/ (Cited on page 49).

[91] Maxime Delmas *et al.* "FORUM: building a Knowledge Graph from public databases and scientific literature to extract associations between chemicals and diseases". In: *Bioinformatics* 37.21 (Nov. 2021), pp. 3896–3904. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btab627. URL: https://doi.org/10.1093/bioinformatics/btab627 (Cited on pages 49, 65).

[92] Michael Ashburner *et al.* "Gene Ontology: tool for the unification of biology". en. In: *Nature Genetics* 25.1 (May 2000). Number: 1 Publisher: Nature Publishing Group, pp. 25–29. ISSN: 1546-1718. DOI: 10.1038/75556. URL: https://www.nature.com/articles/ng0500_25 (Cited on pages 50, 142).

[93] The Gene Ontology Consortium *et al.* "The Gene Ontology knowledgebase in 2023". In: *Genetics* 224.1 (May 2023), iyad031. ISSN: 1943-2631. DOI: 10.1093/genetics/iyad031. URL: https://doi.org/10.1093/genetics/iyad031 (Cited on pages 50, 142).

[94] Paul Shannon *et al.* "Cytoscape: a software environment for integrated models of biomolecular interaction networks". eng. In: *Genome Research* 13.11 (Nov. 2003), pp. 2498–2504. ISSN: 1088-9051. DOI: 10.1101/gr.1239303 (Cited on page 51).

[95] Ludovic Cottret *et al.* "MetExplore: collaborative edition and exploration of metabolic networks". In: *Nucleic Acids Research* 46.Web Server issue (July 2018), W495–W502. ISSN: 0305-1048. DOI: 10.1093/nar/gky301. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6030842/ (Cited on page 51).

[96] Maxime Chazalviel *et al.* "MetExploreViz: web component for interactive metabolic network visualization". eng. In: *Bioinformatics (Oxford, England)* 34.2 (Jan. 2018), pp. 312–313. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btx588 (Cited on page 51).

[97] Vincent Lacroix *et al.* "An introduction to metabolic networks and their structural analysis". eng. In: *IEEE/ACM transactions on computational biology and bioinformatics* 5.4 (2008), pp. 594–617. ISSN: 1557-9964. DOI: 10.1109/TCBB.2008.79 (Cited on page 51).

[98] P Mendes and D Kell. "Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation." In: *Bioinformatics* 14.10 (Jan. 1998), pp. 869–883. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/14.10.869. URL: https://doi.org/10.1093/bioinformatics/14.10.869 (Cited on page 53).

[99] Carmen G. Moles, Pedro Mendes, and Julio R. Banga. "Parameter Estimation in Biochemical Pathways: A Comparison of Global Optimization Methods". en. In: *Genome Research* 13.11 (Nov. 2003). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 2467–2474. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.1262503. URL: https://genome.cshlp.org/content/13/11/2467 (Cited on page 53).

[100] Ralf Steuer *et al.* "Structural kinetic modeling of metabolic networks". In: *Proceedings of the National Academy of Sciences* 103.32 (Aug. 2006). Publisher: Proceedings of the National Academy of Sciences, pp. 11868–11873. DOI: 10.1073/pnas.0600013103. URL: https://www.pnas.org/doi/abs/10.1073/pnas.0600013103 (Cited on page 53).

[101] Gengjie Jia, Gregory Stephanopoulos, and Rudiyanto Gunawan. "Incremental parameter estimation of kinetic metabolic network models". In: *BMC Systems Biology* 6.1 (Nov. 2012), p. 142. ISSN: 1752-0509. DOI: 10.1186/1752-0509-6-142. URL: https://doi.org/10.1186/1752-0509-6-142 (Cited on page 53).

[102] Joseph J. Heijnen and Peter J. T. Verheijen. "Parameter identification of in vivo kinetic models: Limitations and challenges". en. In: *Biotechnology Journal* 8.7 (2013). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/biot.2013001 pp. 768–775. ISSN: 1860-7314. DOI: 10.1002/biot.201300105. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/biot.201300105 (Cited on page 53).

[103] Jonathan Strutz *et al.* "Metabolic kinetic modeling provides insight into complex biological questions, but hurdles remain". In: *Current opinion in biotechnology* 59 (Oct. 2019), pp. 24–30. ISSN: 0958-1669. DOI: 10.1016/j.copbio.2019.02.005. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6731160/ (Cited on page 54).

[104] Jeffrey D. Orth, Ines Thiele, and Bernhard Ø Palsson. "What is flux balance analysis?" en. In: *Nature Biotechnology* 28.3 (Mar. 2010). Number: 3 Publisher: Nature Publishing Group, pp. 245–248. ISSN: 1546-1696. DOI: 10.1038/nbt.1614. URL: https://www.nature.com/articles/nbt.1614 (Cited on pages 54, 75).

[105] Ehsan Motamedian *et al.* "TRFBA: an algorithm to integrate genome-scale metabolic and transcriptional regulatory networks with incorporation of expression data". eng. In: *Bioinformatics (Oxford, England)* 33.7 (Apr. 2017), pp. 1057–1063. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btw772 (Cited on page 55).

[106] Filmon Eyassu and Claudio Angione. "Modelling pyruvate dehydrogenase under hypoxia and its role in cancer metabolism". eng. In: *Royal Society Open Science* 4.10 (Oct. 2017), p. 170360. ISSN: 2054-5703. DOI: 10.1098/rsos.170360 (Cited on page 55).

[107] Pablo Rodríguez-Mier *et al. DEXOM: Diversity-based enumeration of optimal context-specific metabolic networks*. en. Pages: 2020.07.17.208918 Section: New Results. July 2020. DOI: 10.1101/2020.07.17.208918. URL:

https://www.biorxiv.org/content/10.1101/2020.07.17.208918v1 (Cited on pages 55, 173).

[108] Hadas Zur, Eytan Ruppin, and Tomer Shlomi. "iMAT: an integrative metabolic analysis tool". In: *Bioinformatics* 26.24 (Dec. 2010), pp. 3140–3142. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq602. URL: https://doi.org/10.1093/bioinformatics/btq602 (Cited on page 55).

[109] Claudio Angione. "Human Systems Biology and Metabolic Modelling: A Review—From Disease Metabolism to Precision Medicine". In: *BioMed Research International* 2019 (June 2019), p. 8304260. ISSN: 2314-6133. DOI: 10.1155/2019/8304260. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6590590/ (Cited on page 55).

[110] Aarash Bordbar, Neema Jamshidi, and Bernhard O Palsson. "iAB-RBC-283: A proteomically derived knowledge-base of erythrocyte metabolism that can be used to simulate its physiological and patho-physiological states". In: *BMC Systems Biology* 5 (July 2011), p. 110. ISSN: 1752-0509. DOI: 10.1186/1752-0509-5-110. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3158119/ (Cited on page 55).

[111] Ines Thiele *et al.* "Personalized whole-body models integrate metabolism, physiology, and the gut microbiome". In: *Molecular Systems Biology* 16.5 (May 2020). Publisher: John Wiley & Sons, Ltd, e8982. ISSN: 1744-4292. DOI: 10.15252/msb.20198982. URL: https://www.embopress.org/doi/full/10.15252/msb.20198982 (Cited on pages 55, 177).

[112] François-Olivier Desmet *et al.* "Human Splicing Finder: an online bioinformatics tool to predict splicing signals". In: *Nucleic Acids Research* 37.9 (May 2009), e67. ISSN: 0305-1048. DOI: 10.1093/nar/gkp215. URL: https://doi.org/10.1093/nar/gkp215 (Cited on page 56).

[113] Magnus Wang and Antonio Marín. "Characterization and prediction of alternative splice sites". In: *Gene* 366.2 (Feb. 2006), pp. 219–227. ISSN: 0378-1119. DOI: 10.1016/j.gene.2005.07.015. URL: https://www.sciencedirect.com/science/article/pii/S0378111905004117 (Cited on page 56).

[114] Qi Liu, Leiming Fang, and Chengjun Wu. "Alternative Splicing and Isoforms: From Mechanisms to Diseases". en. In: *Genes* 13.3 (Mar. 2022). Number: 3 Publisher: Multidisciplinary Digital Publishing Institute, p. 401. ISSN: 2073-4425. DOI: 10.3390/genes13030401. URL: https://www.mdpi.com/2073-4425/13/3/401 (Cited on page 56).

[115] D. Blangy, H. Buc, and J. Monod. "Kinetics of the allosteric interactions of phosphofructokinase from Escherichia coli". In: *Journal of Molecular Biology* 31.1 (Jan. 1968), pp. 13–35. ISSN: 0022-2836. DOI: 10.1016/0022-2836(68)90051-X. URL: https://www.sciencedirect.com/science/article/pii/002228366890051X (Cited on page 57).

[116] Francisco Llaneras and Jesús Picó. "Stoichiometric modelling of cell metabolism". In: *Journal of Bioscience and Bioengineering* 105.1 (Jan. 2008), pp. 1–11. ISSN: 1389-1723. DOI: 10.1263/jbb.105.1. URL: https://www.sciencedirect.com/science/article/pii/S1389172308700173 (Cited on page 59).

[117] Meghna Rajvanshi and Kareenhalli V. Venkatesh. "Flux Balance Analysis". en. In: *Encyclopedia of Systems Biology*. Ed. by Werner Dubitzky *et al.* New York, NY: Springer, 2013, pp. 749–752. ISBN: 978-1-4419-9863-7. DOI: 10.1007/978-1-4419-9863-7_1085. URL: https://doi.org/10.1007/978-1-4419-9863-7_1085 (Cited on page 60).

[118] Sharon J. Wiback *et al.* "Monte Carlo sampling can be used to determine the size and shape of the steady-state flux space". en. In: *Journal of Theoretical Biology* 228.4 (June 2004), pp. 437–447. ISSN: 0022-5193. DOI:

10.1016/j.jtbi.2004.02.006. URL: https://www.sciencedirect.com/science/article/pii/S0022519304000554 (Cited on page 62).

[119] Ines Thiele *et al.* "Candidate Metabolic Network States in Human Mitochondria: IMPACT OF DIABETES, ISCHEMIA, AND DIET *". English. In: *Journal of Biological Chemistry* 280.12 (Mar. 2005). Publisher: Elsevier, pp. 11683–11695. ISSN: 0021-9258, 1083-351X. DOI: 10.1074/jbc.M409072200. URL: https://www.jbc.org/article/S0021-9258(20)80876-X/abstract (Cited on page 62).

[120] Maike K. Aurich *et al.* "Prediction of intracellular metabolic states from extracellular metabolomic data". en. In: *Metabolomics* 11.3 (June 2015), pp. 603–619. ISSN: 1573-3890. DOI: 10.1007/s11306-014-0721-3. URL: https://doi.org/10.1007/s11306-014-0721-3 (Cited on page 64).

[121] Monica L. Mo, Bernhard Ø Palsson, and Markus J. Herrgård. "Connecting extracellular metabolomic measurements to intracellular flux states in yeast". In: *BMC Systems Biology* 3.1 (Mar. 2009), p. 37. ISSN: 1752-0509. DOI: 10.1186/1752-0509-3-37. URL: https://doi.org/10.1186/1752-0509-3-37 (Cited on pages 64, 83).

[122] Chaitra Sarathy *et al.* "Comparison of metabolic states using genome-scale metabolic models". en. In: *PLOS Computational Biology* 17.11 (Nov. 2021). Publisher: Public Library of Science, e1009522. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1009522. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1009522 (Cited on page 64).

[123] A. W. L. Bayci *et al.* "Metabolomic identification of diagnostic serum-based biomarkers for advanced stage melanoma". eng. In: *Metabolomics: Official Journal of the Metabolomic Society* 14.8 (Aug. 2018), p. 105. ISSN: 1573-3890. DOI: 10.1007/s11306-018-1398-9 (Cited on page 65).

[124] J. Wang *et al.* "Discovery of potential biomarkers for osteoporosis using LC-MS/MS metabolomic methods". eng. In: *Osteoporosis international: a journal established as result of cooperation between the European Foundation for Osteoporosis and the National Osteoporosis Foundation of the USA* 30.7 (July 2019), pp. 1491–1499. ISSN: 1433-2965. DOI: 10.1007/s00198-019-04892-0 (Cited on page 65).

[125] Alexandra Contreras-Jodar *et al.* "Heat stress modifies the lactational performances and the urinary metabolomic profile related to gastrointestinal microbiota of dairy goats". In: *PLoS ONE* 14.2 (Feb. 2019), e0202457. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0202457. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6368375/ (Cited on page 65).

[126] Adam Amara *et al.* "Networks and Graphs Discovery in Metabolomics Data Analysis and Interpretation". In: *Frontiers in Molecular Biosciences* 9 (2022). ISSN: 2296-889X. URL: https://www.frontiersin.org/articles/10.3389/fmolb.2022.841373 (Cited on page 65).

[127] Clément Frainay *et al.* "MetaboRank: network-based recommendation system to interpret and enrich metabolomics results". eng. In: *Bioinformatics (Oxford, England)* 35.2 (Jan. 2019), pp. 274–283. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/bty577 (Cited on page 65).

[128] Tomer Shlomi, Moran N. Cabili, and Eytan Ruppin. "Predicting metabolic biomarkers of human inborn errors of metabolism". eng. In: *Molecular Systems Biology* 5 (2009), p. 263. ISSN: 1744-4292. DOI: 10.1038/msb.2009.22 (Cited on pages 66, 96, 107, 108, 110, 167, 168).

[129] Thierry D.G.A. Mondeel, Vivian Ogundipe, and Hans V. Westerhoff. "Replication of T. Shlomi, M.N. Cabili, E. Ruppin (2009) "Predicting metabolic biomarkers of human inborn errors of metabolism"". In: (May

2018). DOI: 10.5281/zenodo.1254630. URL: https://zenodo.org/record/1254630 (Cited on page 66).

[130] Ali Ebrahim *et al.* "COBRApy: COnstraints-Based Reconstruction and Analysis for Python". In: *BMC Systems Biology* 7.1 (Aug. 2013), p. 74. ISSN: 1752-0509. DOI: 10.1186/1752-0509-7-74. URL: https://doi.org/10.1186/1752-0509-7-74 (Cited on pages 66, 69, 93, 101).

[131] Swagatika Sahoo *et al.* "A compendium of inborn errors of metabolism mapped onto the human metabolic network". en. In: *Molecular BioSystems* 8.10 (Aug. 2012). Publisher: The Royal Society of Chemistry, pp. 2545–2558. ISSN: 1742-2051. DOI: 10.1039/C2MB25075F. URL: https://pubs.rsc.org/en/content/articlelanding/2012/mb/c2mb25075f (Cited on pages 67, 98, 99, 101–103, 108).

[132] The MathWorks Inc. *MATLAB version: 9.13.0 (R2022b)*. Natick, Massachusetts, United States, 2022. URL: https://www.mathworks.com (Cited on page 69).

[133] Laurent Heirendt *et al.* "Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0". en. In: *Nature Protocols* 14.3 (Mar. 2019). Number: 3 Publisher: Nature Publishing Group, pp. 639–702. ISSN: 1750-2799. DOI: 10.1038/s41596-018-0098-2. URL: https://www.nature.com/articles/s41596-018-0098-2 (Cited on pages 69, 82).

[134] Jan Schellenberger *et al.* "Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0". en. In: *Nature Protocols* 6.9 (Sept. 2011). Number: 9 Publisher: Nature Publishing Group, pp. 1290–1307. ISSN: 1750-2799. DOI: 10.1038/nprot.2011.308. URL: https://www.nature.com/articles/nprot.2011.308 (Cited on page 69).

[135] Sjoerd Opdam *et al.* "A systematic evaluation of methods for tailoring genome-scale metabolic models". In: *Cell systems* 4.3 (Mar. 2017), 318–329.e6. ISSN: 2405-4712. DOI: 10.1016/j.cels.2017.01.010. URL:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5526624/ (Cited on page 76).

[136] Daniel Machado and Markus Herrgård. "Systematic Evaluation of Methods for Integration of Transcriptomic Data into Constraint-Based Models of Metabolism". en. In: *PLOS Computational Biology* 10.4 (Apr. 2014). Publisher: Public Library of Science, e1003580. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1003580. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003580 (Cited on page 76).

[137] Robert L. Smith. "Efficient Monte Carlo Procedures for Generating Points Uniformly Distributed over Bounded Regions". In: *Operations Research* 32.6 (Dec. 1984). Publisher: INFORMS, pp. 1296–1308. ISSN: 0030-364X. DOI: 10.1287/opre.32.6.1296. URL: https://pubsonline.informs.org/doi/10.1287/opre.32.6.1296 (Cited on page 82).

[138] Jan Schellenberger and Bernhard Ø Palsson. "Use of Randomized Sampling for Analysis of Metabolic Networks *". English. In: *Journal of Biological Chemistry* 284.9 (Feb. 2009). Publisher: Elsevier, pp. 5457–5461. ISSN: 0021-9258, 1083-351X. DOI: 10.1074/jbc.R800048200. URL: https://www.jbc.org/article/S0021-9258(20)70868-9/abstract (Cited on page 83).

[139] Wout Megchelenbrink, Martijn Huynen, and Elena Marchiori. "optGpSampler: An Improved Tool for Uniformly Sampling the Solution-Space of Genome-Scale Metabolic Networks". en. In: *PLOS ONE* 9.2 (Feb. 2014). Publisher: Public Library of Science, e86587. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0086587. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0086587 (Cited on pages 83, 93).

[140] C. E. Dent and G. R. Philpot. "Xanthinuria, an inborn error (or deviation) of metabolism". eng. In: *Lancet (London, England)* 266.6804 (Jan. 1954), pp. 182–185. ISSN: 0140-6736. DOI: 10.1016/s0140-6736(54)91257-x (Cited on page 106).

[141] Nina Arikyants *et al.* "Xanthinuria type I: a rare cause of urolithiasis". en. In: *Pediatric Nephrology* 22.2 (Feb. 2007), pp. 310–314. ISSN: 1432-198X. DOI: 10.1007/s00467-006-0267-3. URL: https://doi.org/10.1007/s00467-006-0267-3 (Cited on page 107).

[142] Karsten Suhre *et al.* "Human metabolic individuality in biomedical and pharmaceutical research". en. In: *Nature* 477.7362 (Sept. 2011). Number: 7362 Publisher: Nature Publishing Group, pp. 54–60. ISSN: 1476-4687. DOI: 10.1038/nature10354. URL: https://www.nature.com/articles/nature10354 (Cited on page 110).

[143] Thomas Illig *et al.* "A genome-wide perspective of genetic variation in human metabolism". en. In: *Nature Genetics* 42.2 (Feb. 2010). Number: 2 Publisher: Nature Publishing Group, pp. 137–141. ISSN: 1546-1718. DOI: 10.1038/ng.507. URL: https://www.nature.com/articles/ng.507 (Cited on page 111).

[144] William McLaren *et al.* "The Ensembl Variant Effect Predictor". In: *Genome Biology* 17.1 (June 2016), p. 122. ISSN: 1474-760X. DOI: 10.1186/s13059-016-0974-4. URL: https://doi.org/10.1186/s13059-016-0974-4 (Cited on page 112).

[145] Nicola Longo, Marta Frigeni, and Marzia Pasquali. "CARNITINE TRANSPORT AND FATTY ACID OXIDATION". In: *Biochimica et biophysica acta* 1863.10 (Oct. 2016), pp. 2422–2435. ISSN: 0006-3002. DOI: 10.1016/j.bbamcr.2016.01.023. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4967041/ (Cited on page 118).

[146] G. Shilpa Reddy and M. Sujatha. "A Rare Case of Short-Chain Acyl-COA Dehydrogenase Deficiency: The Apparent Rarity of the Disorder Results in Under Diagnosis". In: *Indian Journal of Clinical Biochemistry* 26.3 (July 2011), pp. 312–315. ISSN: 0970-1915. DOI: 10.1007/s12291-011-0139-x. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3162956/ (Cited on page 119).

[147] Nidhi Rani *et al.* "Functional annotation of putative fadE9 of Mycobacterium tuberculosis as isobutyryl-CoA dehydrogenase involved in valine catabolism". eng. In: *International Journal of Biological Macromolecules* 122 (Feb. 2019), pp. 45–57. ISSN: 1879-0003. DOI: 10.1016/j.ijbiomac.2018.10.040 (Cited on page 119).

[148] C. R. Roe *et al.* "Isolated isobutyryl-CoA dehydrogenase deficiency: an unrecognized defect in human valine metabolism". eng. In: *Molecular Genetics and Metabolism* 65.4 (Dec. 1998), pp. 264–271. ISSN: 1096-7192. DOI: 10.1006/mgme.1998.2758 (Cited on page 119).

[149] Pablo Moreno *et al.* "BiNChE: A web tool and library for chemical enrichment analysis based on the ChEBI ontology". In: *BMC Bioinformatics* 16.1 (Feb. 2015), p. 56. ISSN: 1471-2105. DOI: 10.1186/s12859-015-0486-3. URL: https://doi.org/10.1186/s12859-015-0486-3 (Cited on page 122).

[150] Ayşe Demirkan *et al.* "Genome-Wide Association Study Identifies Novel Loci Associated with Circulating Phospho- and Sphingolipid Concentrations". en. In: *PLoS Genetics* 8.2 (Feb. 2012). Publisher: PLOS. DOI: 10.1371/journal.pgen.1002490. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3280968/ (Cited on page 123).

[151] Chad M. Paton and James M. Ntambi. "Biochemical and physiological function of stearoyl-CoA desaturase". In: *American Journal of Physiology - Endocrinology and Metabolism* 297.1 (July 2009), E28–E37. ISSN: 0193-1849.

DOI: 10.1152/ajpendo.90897.2008. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2711665/ (Cited on page 123).

[152] Benjamin Jenkins, James A. West, and Albert Koulman. "A Review of Odd-Chain Fatty Acid Metabolism and the Role of Pentadecanoic Acid (C15:0) and Heptadecanoic Acid (C17:0) in Health and Disease". In: *Molecules* 20.2 (Jan. 2015), pp. 2425–2444. ISSN: 1420-3049. DOI: 10.3390/molecules20022425. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6272531/ (Cited on page 124).

[153] Dinesh Kumar Barupal and Oliver Fiehn. "Chemical Similarity Enrichment Analysis (ChemRICH) as alternative to biochemical pathway mapping for metabolomic datasets". en. In: *Scientific Reports* 7.1 (Nov. 2017). Number: 1 Publisher: Nature Publishing Group, p. 14567. ISSN: 2045-2322. DOI: 10.1038/s41598-017-15231-w. URL: https://www.nature.com/articles/s41598-017-15231-w (Cited on page 124).

[154] Jana Sajovic *et al.* "The Role of Vitamin A in Retinal Diseases". In: *International Journal of Molecular Sciences* 23.3 (Jan. 2022), p. 1014. ISSN: 1422-0067. DOI: 10.3390/ijms23031014. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8835581/ (Cited on page 130).

[155] Lin Wang *et al.* "A review of computational tools for design and reconstruction of metabolic pathways". In: *Synthetic and Systems Biotechnology* 2.4 (Dec. 2017), pp. 243–252. ISSN: 2405-805X. DOI: 10.1016/j.synbio.2017.11.002. URL: https://www.sciencedirect.com/science/article/pii/S2405805X17300820 (Cited on page 147).

[156] Hulda S Haraldsdóttir *et al.* "CHRR: coordinate hit-and-run with rounding for uniform sampling of constraint-based models". In: *Bioinformatics* 33.11 (June 2017), pp. 1741–1743. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx052. URL: https://doi.org/10.1093/bioinformatics/btx052 (Cited on page 167).

[157] Johann F. Jadebeck, Wolfgang Wiechert, and Katharina Nöh. *CHRRT: boosting coordinate hit-and-run with rounding by thinning*. en. Pages: 2022.11.17.516802 Section: New Results. Nov. 2022. DOI: 10.1101/2022.11.17.516802. URL: https://www.biorxiv.org/content/10.1101/2022.11.17.516802v1 (Cited on page 167).

[158] Axel Theorell *et al.* "PolyRound: polytope rounding for random sampling in metabolic networks". In: *Bioinformatics* 38.2 (Jan. 2022), pp. 566–567. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btab552. URL: https://doi.org/10.1093/bioinformatics/btab552 (Cited on page 167).

[159] Colin A. Smith *et al.* "XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification". eng. In: *Analytical Chemistry* 78.3 (Feb. 2006), pp. 779–787. ISSN: 0003-2700. DOI: 10.1021/ac051437y (Cited on page 169).

[160] Mark D. Wilkinson *et al.* "The FAIR Guiding Principles for scientific data management and stewardship". en. In: *Scientific Data* 3.1 (Mar. 2016). Number: 1 Publisher: Nature Publishing Group, p. 160018. ISSN: 2052-4463. DOI: 10.1038/sdata.2016.18. URL: https://www.nature.com/articles/sdata201618 (Cited on page 171).

[161] Nathalie Poupin *et al.* "Improving lipid mapping in Genome Scale Metabolic Networks using ontologies". eng. In: *Metabolomics: Official Journal of the Metabolomic Society* 16.4 (Mar. 2020), p. 44. ISSN: 1573-3890. DOI: 10.1007/s11306-020-01663-5 (Cited on page 172).

[162] Markus W. Covert and Bernhard Ø Palsson. "Transcriptional Regulation in Constraints-based Metabolic Models of Escherichia coli * 210". English. In: *Journal of Biological Chemistry* 277.31 (Aug. 2002). Publisher: Elsevier, pp. 28058–28064. ISSN: 0021-9258, 1083-351X. DOI: 10.1074/jbc.M201691200. URL: https://www.jbc.org/article/S0021-9258(19)66275-7/abstract (Cited on page 172).

[163] Javier Carrera *et al.* "An integrative, multi-scale, genome-wide model reveals the phenotypic landscape of Escherichia coli". In: *Molecular Systems Biology* 10.7 (July 2014). Publisher: John Wiley & Sons, Ltd, p. 735. ISSN: 1744-4292. DOI: 10.15252/msb.20145108. URL: https://www.embopress.org/doi/full/10.15252/msb.20145108 (Cited on page 172).

[164] José P. Faria *et al.* "Genome-scale bacterial transcriptional regulatory networks: reconstruction and integrated analysis with metabolic models". In: *Briefings in Bioinformatics* 15.4 (July 2014), pp. 592–611. ISSN: 1467-5463. DOI: 10.1093/bib/bbs071. URL: https://doi.org/10.1093/bib/bbs071 (Cited on page 172).

[165] Hasan Özen. "Glycogen storage diseases: New perspectives". In: *World Journal of Gastroenterology : WJG* 13.18 (May 2007), pp. 2541–2553. ISSN: 1007-9327. DOI: 10.3748/wjg.v13.i18.2541. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4146814/ (Cited on page 173).

[166] Tomer Shlomi *et al.* "Network-based prediction of human tissue-specific metabolism". en. In: *Nature Biotechnology* 26.9 (Sept. 2008). Number: 9 Publisher: Nature Publishing Group, pp. 1003–1010. ISSN: 1546-1696. DOI: 10.1038/nbt.1487. URL: https://www.nature.com/articles/nbt.1487 (Cited on page 173).

[167] Yuanyuan Xiang *et al.* "Recreational Nitrous Oxide Abuse: Prevalence, Neurotoxicity, and Treatment". en. In: *Neurotoxicity Research* 39.3 (June 2021), pp. 975–985. ISSN: 1476-3524. DOI: 10.1007/s12640-021-00352-y. URL: https://doi.org/10.1007/s12640-021-00352-y (Cited on page 175).

[168] H. Kondo *et al.* "Nitrous oxide has multiple deleterious effects on cobalamin metabolism and causes decreases in activities of both mammalian cobalamin-dependent enzymes in rats". eng. In: *The Journal*

of Clinical Investigation 67.5 (May 1981), pp. 1270–1283. ISSN: 0021-9738. DOI: 10.1172/jci110155 (Cited on page 175).

[169] G. Grzych *et al.* "L'acide méthylmalonique : un marqueur spécifique de l'intoxication chronique au protoxyde d'azote ?" In: *La Revue de Médecine Interne* 43.3 (Mar. 2022), pp. 197–198. ISSN: 0248-8663. DOI: 10.1016/j.revmed.2022.01.001. URL: https://www.sciencedirect.com/science/article/pii/S0248866322000017 (Cited on page 176).

[170] Andrew J. Waclawik *et al.* "Myeloneuropathy from nitrous oxide abuse: unusually high methylmalonic acid and homocysteine levels". eng. In: *WMJ: official publication of the State Medical Society of Wisconsin* 102.4 (2003), pp. 43–45. ISSN: 1098-1861 (Cited on page 176).

[171] Rosemary Deacon *et al.* "SELECTIVE INACTIVATION OF VITAMIN B12 IN RATS BY NITROUS OXIDE". In: *The Lancet*. Originally published as Volume 2, Issue 8098 312.8098 (Nov. 1978), pp. 1023–1024. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(78)92341-3. URL: https://www.sciencedirect.com/science/article/pii/S0140673678923413 (Cited on page 177).

[172] Sheng Hui *et al.* "Quantitative Fluxomics of Circulating Metabolites". eng. In: *Cell Metabolism* 32.4 (Oct. 2020), 676–688.e4. ISSN: 1932-7420. DOI: 10.1016/j.cmet.2020.07.013 (Cited on page 177).

[173] Philippa D. Darbre. "Endocrine Disruptors and Obesity". In: *Current Obesity Reports* 6.1 (2017), pp. 18–27. ISSN: 2162-4968. DOI: 10.1007/s13679-017-0240-4. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5359373/ (Cited on page 179).

[174] Elizabeth M. Martin, Miroslav Stýblo, and Rebecca C. Fry. "Genetic and epigenetic mechanisms underlying arsenic-associated diabetes mellitus: a perspective of the current evidence". en. In: *Epigenomics* 9.5 (May 2017). Publisher: Future Science Group, p. 701. DOI: 10.2217/epi-2016-0097.

URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5480787/ (Cited on page 179).

[175] Philippe Grandjean *et al.* "Estimated exposures to perfluorinated compounds in infancy predict attenuated vaccine antibody concentrations at age 5-years". eng. In: *Journal of Immunotoxicology* 14.1 (Dec. 2017), pp. 188–195. ISSN: 1547-6901. DOI: 10.1080/1547691X.2017.1360968 (Cited on page 179).

[176] Nadeem Ghani Khan *et al.* "A comprehensive review on the carcinogenic potential of bisphenol A: clues and evidence". In: *Environmental Science and Pollution Research International* 28.16 (2021), pp. 19643–19663. ISSN: 0944-1344. DOI: 10.1007/s11356-021-13071-w. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8099816/ (Cited on page 179).

[177] Laura N. Vandenberg *et al.* "Urinary, Circulating, and Tissue Biomonitoring Studies Indicate Widespread Exposure to Bisphenol A". In: *Environmental Health Perspectives* 118.8 (Aug. 2010), pp. 1055–1070. ISSN: 0091-6765. DOI: 10.1289/ehp.0901716. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2920080/ (Cited on page 179).

[178] Beverly S. Rubin. "Bisphenol A: An endocrine disruptor with widespread exposure and multiple effects". In: *The Journal of Steroid Biochemistry and Molecular Biology*. Endocrine Disruptors 127.1 (Oct. 2011), pp. 27–34. ISSN: 0960-0760. DOI: 10.1016/j.jsbmb.2011.05.002. URL: https://www.sciencedirect.com/science/article/pii/S0960076011001063 (Cited on page 179).

[179] Donald G. Robertson, Paul B. Watkins, and Michael D. Reily. "Metabolomics in Toxicology: Preclinical and Clinical Applications". In: *Toxicological Sciences* 120.suppl_1 (Mar. 2011), S146–S170. ISSN: 1096-6080. DOI: 10.

1093/toxsci/kfq358. URL: https://doi.org/10.1093/toxsci/kfq358 (Cited on page 179).

[180] Aurélie Roux *et al.* "Applications of liquid chromatography coupled to mass spectrometry-based metabolomics in clinical chemistry and toxicology: A review". In: *Clinical Biochemistry*. Mass Spectrometry in Laboratory Medicine 44.1 (Jan. 2011), pp. 119–135. ISSN: 0009-9120. DOI: 10.1016/j.clinbiochem.2010.08.016. URL: https://www.sciencedirect.com/science/article/pii/S0009912010003759 (Cited on page 179).