Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Biotechnology and Food Science

Helge Monsson

Innovative Exploration of the Role of Biomass Composition and Growth Rate on Metabolic Flux Predictions in Genome-Scale Models

Master's thesis in Biotechnology
Supervisor: Eivind Almaas
Co-supervisor: Vetle Simensen

May 2023

NTNU
Norwegian University of
Science and Technology

Helge Monsson

Innovative Exploration of the Role of Biomass Composition and Growth Rate on Metabolic Flux Predictions in Genome-Scale Models

Master's thesis in Biotechnology
Supervisor: Eivind Almaas
Co-supervisor: Vetle Simensen
May 2023

Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Biotechnology and Food Science

NTNU
Norwegian University of
Science and Technology

Acknowledgements

First I would like to extend my thanks to the Network Systems Biology group for making this last year a wonderful one. The general motivation and passion for systems biology displayed by all the members of the group helped cultivate my own interest for the subject. I would like to especially thank Vetle Simensen for suffering my questions and general help at steering the project. The feedback he provided on my ideas and the assistance he provided were invaluable for this project. I would also like to thank my professor Eivind Almaas for allowing me this opportunity to write a project under his research group. The environment he has cultivated and the level of depth in his knowledge has served as a major motivation for this project.

Furthermore I would like to thank my fellow study mates, Daniel, Lars and Sigurd for keeping me motivated and making these two years in biotechnology a fantastic experience. A special thanks also goes out to my friend Mike Bulte, who provided invaluable guidance for the coding parts of this project.

Finally I would like to thank my family and my father who nurtured a deep passion for research and knowledge within me, which has driven me these past 7 years.

Abstract

The collection of biological data in the form of large data sets has been an undertaking in the field of biology for the last few decades. Due to the nature of biological systems these data sets are large and complex. A field that seeks to make this data clearer and more understandable is the field of systems biology, which utilises a holistic approach to biological processes, understanding them as interconnected systems. One of the sub-fields of systems biology is the field of genome scale metabolic modelling, which utilises network analysis in combination with genomic data in order to develop models for understanding the metabolism of cells. These models, known as Genome Scale Metabolic Models (GEMs), are representative of specific cells and are capable of predicting the activity of pathways in the metabolic network of the cell. These predictions are known as metabolic flux distributions, where flux is a value representing the metabolic flow through a reaction. It calculates these flux predictions by utilising flux balance analysis (FBA) in conjunction with a pseudo-reaction that represents the cells consumption of metabolites for growth. This pseudo-reaction is the biomass objective function (BOF). However, the BOFs for the models are usually made to be static and unchanging with factors like growth rate and environment. This is an assumption that crucially does not reflect reality.

A previous study researched how uncertainty in the biomass composition would impact the predicted flux distributions. However, this study was performed on a model that is naturally robust and therefore is likely to underestimate the impact of the uncertainty injections. The current study performed a similar analysis on a more restricted model with a lower degree of robustness, namely the eciML1515 model of *E. coli*. It then implemented growth rate-specific BOFs for the iML1515, based on experimental data sets, in order to study if the accuracy of the model increases with a more condition-dependent BOF. Results found that the eciML1515 displayed generally high degrees of propagation of uncertainty from the biomass composition to the biomass production, as well as some central metabolic pathways. While both models displayed a level of robustness, the standard GEM was far more robust then the eciML1515 model. These results support the idea that more constrained models, compared to the standard GEMs, are likely to be more impacted by perturbations and alterations of the biomass composition.

As for the growth rate-specific BOFs, they displayed a general improvement of accuracy on the flux distribution predictions for most of the experimental data sets. The exception to this was for the data sets corresponding to acetate, glucose and gluconate as carbon sources. For glucose this result was expected, as the standard BOF is based on specific data gathered from *E. coli* grown on glucose.

In conclusion, this study finds that the uncertainty propagation of the models predicted flux distributions is very likely to increase the more restricted the models are. However, by implementing condition-specific BOFs and accounting for some of the variations caused by environments and growth rate, the accuracy of GEMs can increase. More research should be done into environmental and conditional factors that can be accounted for in GEMs, as well as a closer study on how the biomass composition can impact flux distributions, focusing on specific pathways in the metabolic network.

Sammendrag

Innhentning av biologiske data i form av store datasett har vore eit mål i biologifeltet de siste tiårene. Disse datasetta er store og komplekse, grunnet måten biologiske system har utviklet seg. Eit felt som søker å gjøre disse datasetta klarere og meir forståeleg er systembiologi. Det bruker ein heilskapeleg tilnærming av tilnærming av biologiske prosesser og forstår dem som samankobla system. Ein av underfelta til systembiologi er feltet for genomskala metabolsk modelering, som utnytter nettverksanalyse i kombinasjon med genomdata for å utvikle modeller for å kartlegge metabolismen til celler. Disse modellene, kjent som Genomskala Metabolske Modeller (GEMs), representerer spesifikke celler og er i stand til å predikere aktiviteten til metabolismen i cellen. Disse prediksjonene kalles for metabolske fluksfordelinger, der fluks er ein verdi som representerer den metabolkse strømmen igjennom ein reaksjon. Den beregner disse fluksprediksjonene ved å bruke fluksbalanse analyse (FBA) i kombinasjon med ein pseudoreaksjon som representerer cellens forbruk av metabolitter for vekst. Denne pseudoreaksjonen er biomasseobjektiv funksjonen (BOF). Skjønt BOF-ene til disse modelene er ofte statiske og endrer seg ikkje med faktorer som miljø og vekstrate. Dette er ein antagelse som ikkje gjenspeiler verkelegheiten.

Eit tidligere studie har blitt gjort der dei undersøkte om usikkerhet i den antatte biomasse komposisjonen hadde invirkning på fluksprediksjonene. Dette studiet vart derimot utført på ein model som er naturleg robust og derfor sannsynlegvis undervurderte effekten av usikkerhetsinjeksjonen. Den nåverende studien gjennomførte ein lignende analyse på ein meir begrensa modell med mindre robusthet, i form av eciML1515 til *E. coli*. Videre så implementerte den vekstrate spesifikke BOF-er for iML1515, basert på eksperimentell data, for å undersøke om nøyaktigheten til modellen auker med ein meir spesifikk BOF. Resultata til denne analysen fant at eciML1515 hadde ein høg grad av usikkerhetsspredning fra biomassekomposisjonen til biomasseproduksjon og visse metabolske reaksjonsveier. Begge modellene visste robusthet, men standard GEM hadde langt høgere grad av robusthet samanligna med ecIML1515 modellen. Disse resultata støtter ideen om at meir begrensa modeller, samanligna med standard GEMs, vil sannsynligvis være meir påverka av endringer i biomassekomposisjonen.

Når det gjeld dei vekstrate spesifikke BOF-ene, så viste de generell forbedringer i nøyaktigheten til fluksprediksjonene for de fleste datasetta. Unntaket var for datasetta som tilhørte acetate, glukose og gluconate som kilde til karbon. Det var forventa at standard BOF-en skulle vise meir nøyaktig fluksprediksjon for glukose som karbon kilde da denne BOF-en er basert på data fra ein *E. coli* som brukte glukose som karbon kilde.

Oppsummert, så konkluderer dette studiet med at usikkerhetsspredning for fluksprediksjoner vil mest sannsynleg auke jo meir begrensa modellene er. Dette kan derimot gjørast opp for med å implementere meir spesifikke BOF-er som bedre representerer biomasse komposisjonen under spesifikke forhold. Meir forskning burde utførast på faktorer som kan implementerast i GEMs, i tillegg til eit nærare studie på korleis biomasse komposisjonen påverkar fluksprediksjoner, med fokus på spesifikke reaksjonsveier i metabolismen.

Table of Contents

List of Tables

List of Figures

Abbreviations

GEM	=	Genome-scale Metabolic Model
FBA	=	Flux Balance Analysis
BOF	=	Biomass Objective Function
COBRA	=	Constraint Based Reconstruction and Analysis
LP	=	Linear Programming
pFBA	=	Parsimonious Flux Balance Analysis
E. coli	=	*Escherichia coli*
MOMA	=	Minimisation of Meatbolic Adjustment
GECKO	=	GEM with Enymzatic Constraints using Kinetics and Omics data
GAM	=	Growth Associated Maintenance
NGAM	=	Non Growth Associated Maintenance
MW	=	Molecular Weight
SD	=	Standard Deviation
SDR	=	Standard Deviation Ratio
RMSE	=	Root Mean Square Error
WT	=	Wild Type

Chapter 1

Introduction

The collection of biological data has been a primary goal in the field of biology for well over the last quarter-century, with a significant emphasis on unraveling the genome, proteome and metabolome of various species [1]. The Human-Genome project, which began in the 1990s, took 13 years to map out 90% of the human genome and was a massive endeavour by the scientific field, involving extensive collaboration [2]. Similar projects have been performed on other species with an ever increasing efficiency. Today, it is possible to sequence a genome in less than a week [3], generating complex data sets that explore the biological components of a species in more detail then we've ever been able to before. However, the complexity of these data sets also poses challenges. The interactions between the genome, proteins, and metabolism of a species are intricate and surpass the comprehension of human researchers using only traditional approaches. Consequently, innovations in computation, handling and processing of this data have recently emerged as crucial advances for biology [4]. The idea of using computational tools and approaches to convert these dense and intricate interactions between proteins, genome and metabolism into data fit for interpretation is enabling researchers to keep up with the ever-growing collection of large datasets [5].

Systems biology is an approach that addresses these challenges of managing these large datasets in biology. It is an evolving field that aims to advance our understanding of complex biological systems by analyzing and modeling them as interconnected networks of components. Unlike the more traditional reductionist approach that focuses on individual parts, it has a more holistic approach to the study of biological behaviour, looking at it as an ever-changing system [6]. Systems biology utilises experimental data collected from various fields to map out the interactions between the components of a biological behaviour, before scaling these up to full models describing these behaviours. These models can simulate the system, predict how it will react to changes in its environment and identify key components and pathways [7]. This approach allows for the construction of simplified and more understandable representations of complex biological systems through the use of models. One of the core ways to understand and develop these models is with network analysis. By treating the individual components as nodes and their interactions as edges, the biological behaviour can be understood as a network [8].

The principles of systems biology and network analysis can be applied to metabolic information to create a model of the metabolic network based on genomic data that contains all known metabolic reactions and the corresponding genes. This is done by creating a representation of the cellular components, where the metabolites are nodes and the interactions between them are edges. This is the central principle behind genome scale metabolic modeling, a powerful tool capable of simulating the metabolism of cells, making complex metabolic data easier to understand and interact with. Traditionally, genome scale metabolic models (GEMs) was mainly used to further our understanding of individual organisms. However, in recent times there has been a growing interest in metabolic data for combining it with other data in order to produce a greater understanding of metabolic interactions between organisms. One example is for understanding the interactions between a host biome and its microbial community [9]. GEMs has the potential of being widely adopted to increase our understanding of other fields, increase production of specific drugs by manipulating metabolism of organisms, designing microbial strains optimized for producing specific biomaterials like biofuel, etc [10]. The use of GEMs would allow researchers to understand metabolism in a more comprehensive manner, enabling predictions on how the behaviours of organisms would change based on the environment and how, in turn, the organisms affect the environment. With the increasing availability of genomic and metabolic data the potential applications for GEMs is only growing.

The central prediction produced by GEMs is the metabolic flux distribution of the model. The metabolic fluxes are a value that represents the turnover rate of a metabolite through a reaction. In other words, it creates a mathematical representation of the predicted activity in a metabolic network. To create these predictions it uses mathematical optimization and linear programming, a method for maximising or minimising a vector in an equation set. However, in order for these methods to be properly applicable, an objective function has to be defined.

The objective function, and therefore the target for optimisation in most GEMs, is the biomass objective function (BOF). This is a simplified representation of the organism growth and production of biomass. It is set as the objective function as it is a reasonable assumption that most organisms have evolved to optimize for growth. The principal method of applying the concepts of mathematical optimisation and linear programming to the metabolic network is what is known as flux balance analysis (FBA), which is capable of calculating the optimal configuration for the metabolic fluxes in the network to produce biomass [11]. In order for FBA to represent the metabolism of an organism a couple of key assumptions are made. One of these assumptions is of steady state; namely that the metabolism of the organism has reached a point where the concentration of the metabolic contents in the organism never change, i.e. that the consumption and production of each metabolite internally cancel each other out. A second assumption is that the organism in question has reached a state of optimality. This is done as FBA naturally calculates the most optimal solution for the organism and so would only be an accurate representation if the organism in question has managed to reach the optimal state. This can be a challenge as there might be hindrances stopping the organism from reaching this optimality not accounted for in the metabolic model. A third assumption is the assumption of mass balance, that every reaction in the metabolic network can be properly represented as synthesis or degradation of a set of metabolites, in relatively equal measures. These assumptions are central to FBA and necessary for FBA to keep its elegant and simple use-case.

However, GEMs inherently make another assumption, owing to the static nature of the BOF; namely that the biomass composition of a cell never changes due to environmental and metabolic perturbations. However, studies have shown that the biomass composition of cells can vary with factors like growth rate and environment. As such this assumption is likely incorrect and by factoring in the growth rate for the BOF it is possible that the accuracy of the predicted metabolic network of the model will increase [12].

In spite of these assumptions, GEM has shown itself capable of producing models that show accurate and robust results for predicting the growth rate of organisms. The robustness is likely owing to the flexible nature of GEM and in particular the commonly used model type Constraint

Based Reconstruction and Analysis (COBRA). This model type is defined by a limited number of restrictions. One particular study, by Maranas et al. [13], investigated the propagation of uncertainty through the metabolic network on an *E. coli* GEM. They found that the prediction of growth rate displayed a high level of robustness and resistance to perturbations in the biomass coefficients. However, the prediction of the internal metabolic fluxes did not display this same level of robustness; in some cases even amplifying the uncertainty propagation from the biomass coefficients throughout the metabolic network [13]. This could be a side effect of the flexible nature of the COBRA model, causing huge internal metabolic pathway shifts in order to maintain the same biomass growth in spite of the perturbation. As such a model type that has more restrictions on the internal metabolic network could potentially give a more accurate view on how the metabolic network reacts to alterations in the environment and biomass composition. One such model type is the enzyme constrained models GECKO, which accounts for the reaction enzymes. Enzymes are proteins that assists in the reaction without being a product or substrate [14]. As such the GECKO model is potentially more rigid in its metabolic network predictions and less robust in the biomass growth prediction, but will give a better view of uncertainty propagation throughout the metabolic network.

The overarching aim of this study is to study the metabolic network and how uncertainty and alterations in the biomass composition interacts with the predictions of flux distributions from the model. This overarching aim can be split into two smaller ones.

The first aim is to perform a propagation of uncertainty analysis on the eciML1515, an *E. coli* GECKO model. This is based on the reasoning that the GECKO model will give a more accurate evaluation of predicted flux distributions following perturbations in the biomass composition, owing to the model's lower degree of robustness.

The second aim is to study how a condition dependent biomass representation in the model will affect the accuracy of the flux distributions. This is achieved by creating a new growth rate-specific biomass objective function for an *E. coli* GEM and testing how it performs based on experimental data from Gerosa et al. study [15]. The study collected data on specific reactions, the growth rate and exchange reactions from *E. coli* cells growing on different carbon sources. The principal aim of this study is to further the understanding of predicted internal metabolic reactions, how the biomass composition can affect them and if implementing more specific representations of the biomass can increase the accuracy of the model's predictions.

Chapter 2

Background

This chapter aims to give the reader the necessary understanding of the theoretical foundation behind this project. The chapter starts by presenting an overview of cell metabolism, with a focus on metabolic pathways and enzymes that control the reactions involved with these pathways. Moving forward, the chapter focuses on the representation of the metabolism as genome-scale models. It explores the mathematical and theoretical basis for the process of developing and analyzing these GEM models as well as some of the most common analysis performed on these models such as FBA, Parsimonious Flux Balance Analysis (pFBA) and Minimization of Metabolic Adjustment (MOMA). Finally it gives an overview of two different models that represent *E. coli* in the form of iML1515 and eciML1515.

2.1 Cell Metabolism

The metabolism of a cell can be defined as the complex network of biochemical reactions that enables the cell to sustain life. These reaction includes processes such as energy utilisation, nutrient uptake and building cellular components. Central to the metabolic network are the enzymes, specialized proteins that can act as catalysts to control the biochemical reactions that occur within the cell. They are capable of controlling the energy-requirements and the rate at which internal reactions occur. Enzymes exhibit a high degree of specificity to their specific substrates. A key component is their structure and formation of something called an active site, a pocket or cleft with a structure specific to their respective substrates. Once the substrate binds to the active site, the enzyme is capable of increasing the speed or even inducing the transformation of a substrate into a product. Most enzymes do this by lowering the activation energy, or free energy, needed for the energy to occur by introducing a more stable transition state between substrate and product [14]. Fig. 2.1 shows an example of how an enzyme can change the energy requirements for a reaction to occur. Most enzyme-catalysed reactions are defined by the turnover number, k_{cat}, of the enzyme in question. The formula that defines these turnover numbers is

$$k_{cat} = \frac{V_{max}}{[E]}, \tag{2.1}$$

where V_{max} is the maximum velocity of the biochemical reaction in question and $[E]$ is the concentration of enzymes present. This is a representation of the maximum number of substrate molecules that can be transformed into products per molecule of enzyme per unit time. Due to this it acts as an upper limit on how fast said reaction can occur [16]. Turnover number and its impact on the metabolism of a cell will be discussed further in chapter 2.2.11.

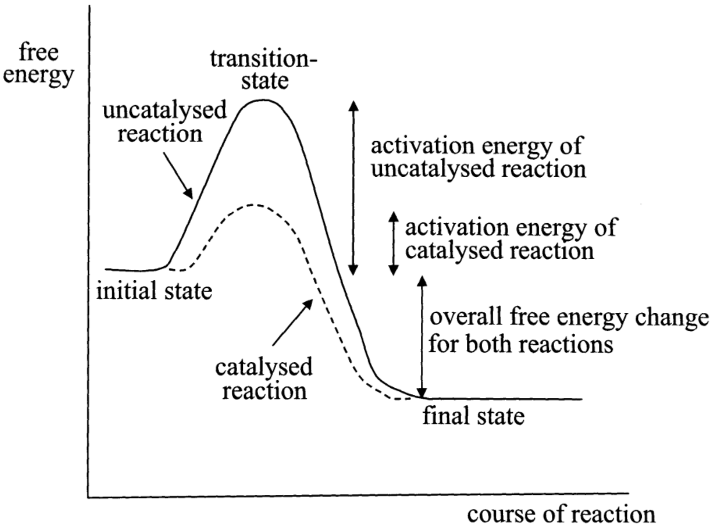

free energy
transition-state
uncatalysed reaction
activation energy of uncatalysed reaction
activation energy of catalysed reaction
initial state
overall free energy change for both reactions
catalysed reaction
final state
course of reaction

Figure 2.1: Graph representing the difference in activation energy required between an uncatalyzed reaction and a catalyzed one. The y-axis represents the free energy required at that moment in the reaction and the x-axis represents the current state of the reaction, moving from the inital state to then final state. The catalyzed reaction has a noticeable lower curve and so a lower required free energy needed to reach the final state. This translates into a higher rate of reaction, or reaction speed. Note that this is a general example and not a specific one. Figure taken from Ref. [14].

The substrate and products of these reactions are referred to as metabolites, which move through different pathways in the cell metabolism. A pathway is a series of interconnected biochemical reactions that convert the substrates into products [17]. Thanks to the myriad of pathways present in the metabolism, cells are capable of turning a relatively limited amount of available extra-cellular substrates into a vast amount of products for different functions and requirements. These pathways can then be used to group and categorize different sections of metabolism, by separating them based on what function they serve in the cell. An example of this is shown in Fig. 2.2 with the metabolism of *E.coli*. While the metabolism might seem like an impossible network to untangle, modern computational methods, a deeper understanding of the genome and a systemic way of thinking has allowed us to make huge strides in representing and understanding these metabolic networks [18].

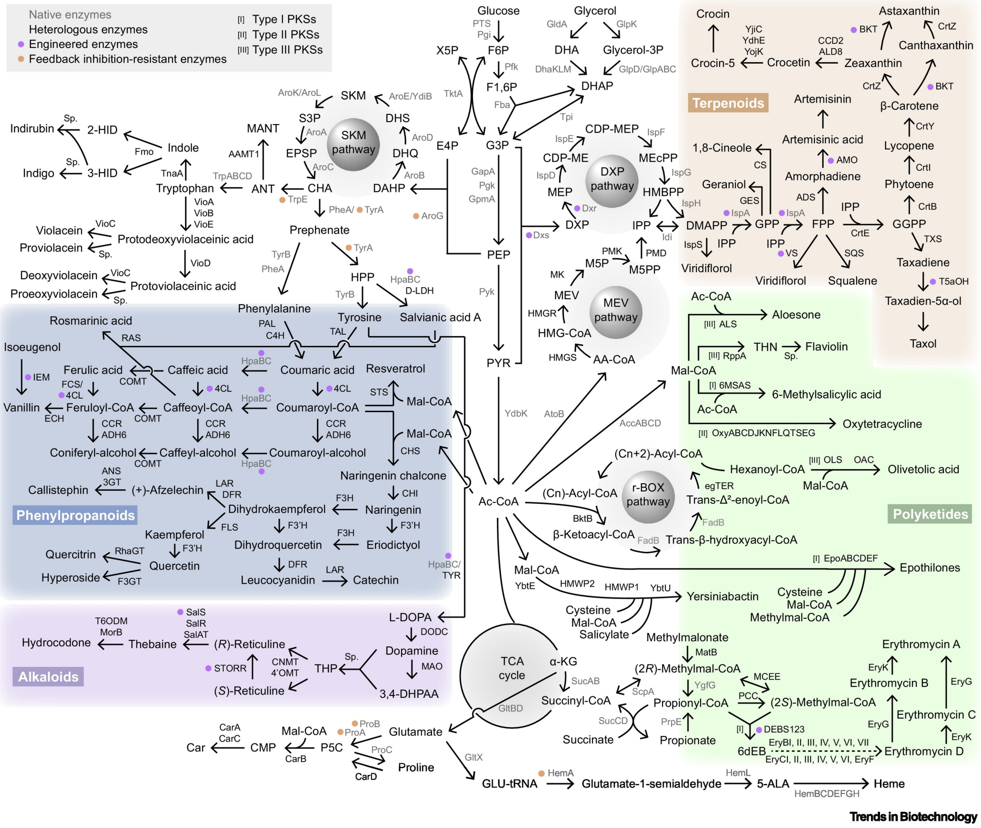


Figure 2.2: Map over the metabolism of an *E.coli* showing some of the different biochemical reactions available to the cell. They are grouped and color coded based on what function they serve for the cell. This shows that, while the metabolism of a cell might seem complex, by looking at the pathways and systems instead of individual reactions the metabolism can be mapped out in a more understandable way. Figure taken from Ref. [19].

2.2 Genome Scale Modeling

Genome-scale modeling is a method for creating a representation of the metabolism of a cell based on genomic data. It is a network based process for collecting all the known metabolic information about a system, including genes, enzymes, metabolites and reactions. It can provide predictions relating to growth and cellular fitness for single cells and allow a deeper understanding of their metabolism [9]. This section will cover the process of creating these GEMs.

2.2.1 Genome Scale Network Reconstruction

The first step in the GEM approach is to create a genome-scale network reconstruction, which is a computational method that translates genomic data into a metabolic network. The starting point is a well-annotated genomic sequence of the target organism, typically obtained from public databases or similarly available data. The annotated genome sequence provides information on what genes are present in the organism and their function. As such it will give an initial set of candidate biochemical reactions that are encoded on the genome, allowing for a draft reconstruction of the network.

This part can often be automated with computer-assisted tools, although the network produced by this is often full of gaps and missing critical reactions or pathways due to limitations of the genome annotation. For example it will be missing specific information about the organism like the substrate and co-factor specificity. To address these gaps and refine the network the next step of the process is to integrate metabolic, genetic and biochemical data to map out which genes and enzymes are involved in which reactions and pathways and fill in the gaps in the draft network produced from the annotated genome sequence. This part can not be fully automated and requires detailed knowledge about the metabolism of the organism in question. The goal is to fill in enough holes in the draft network until a metabolic phenotype is achieved.

The data for these metabolic phenotypes are usually gathered experimentally and gives a suitable target for the network construction. This is a labour intensive and time consuming process. As an example, biochemical studies of the enzymes belonging to the organism can help fill in the gaps in reversibility and substrate specificity that the would not be present in the draft network. Once the gaps have been sufficiently filled, a curated reconstruction of the network has been formed. After this the next step is to turn the curated reconstruction of the network into a mathematical representation. This becomes the basis for the GEM which can then be further developed by adding constraints that limit the metabolism and allow for more biologically realistic predictions. [20].

2.2.2 Constraint Based Reconstruction and Analysis

The conceptual basis behind using constraint based reconstruction and analysis (COBRA) is formed around six axioms that are based on fundamental statements:
Axiom #1: All cellular functions are based on chemistry. As a result of this they can be described by chemical equations.
Axiom #2: Annotated genome sequences combined with experimental data enables a reconstruction of the genome-scale network. This is the basis behind the network reconstruction discussed in the earlier section.
Axiom #3: Cells operate under a variety of constraints. These constraints cannot be violated and is what allows for the estimation of all the states that a genome-scale network reconstruction can achieve. These constraints fall under four different categories that will be discussed further: psycho-chemical, topological, environmental and regulatory.
Axiom #4: Cells function in a context-specific manner. This means that when a cell is placed in a particular environment it expresses a subset of genes in response to said environments cues. In response to this the cellular component of the cell can be profiled with -omics methods (transciptomics, metaboliomics, proteomics etc.) and used to tailor the network to better represent the environment being considered.
Axiom #5: Mass and energy is conserved. This is a fundamental physical law in nature. Due to the nature of this law and that all chemical equations can be represented with stoichiometric coefficients with the set of biochemical equations being represented by the stoichiometric matrix (S), all steady states can be described mathematically by one linear equation, $\mathbf{Sv} = 0$. $\mathbf{v}$ is the vector of fluxes through chemical reactions.
Axiom #6: Cells evolve under selection pressure in a given environment. From this a definition of the objective function can be made as a representation of this selection pressure to determine the optimal states from the given network reconstruction and constraints. This statement will be further explored in subsection 2.2.9.
These six axiomatic statements and the constraints they represent forms the theory behind COBRA which is that a model of the metabolic network can be formed by defining and implement-

ing the constraints into the metabolic network in conjunction with the stoichiometric matrix [21, 22].

2.2.3 The Stoichiometric Matrix

The stoichiometric matrix is an important tool in the context of genome-scale modeling. It is a way to describe the relationship between metabolites and reactions in a metabolic system. The stoichiometric matrix is defined as the matrix S as a $m \times n$ matrix with the form

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m1} & s_{m2} & \cdots & s_{mn} \end{bmatrix}, \tag{2.2}$$

where m represents the metabolites in the system, n represents the reactions and s_{mn} represents the stoichiometric coefficient of the metabolite through the reaction. From this matrix we can define the vector $\mathbf{v}$ given as $n \times 1$ and further define the overall change of metabolites in the system as $\mathbf{Sv}$. From the earlier mentioned mass and energy conservation axiom we can set this value as

$$\mathbf{Sv} = 0. \tag{2.3}$$

This is what is known as the *steady state assumption*, an important assumption in the context of GEMs. By implementing this assumption the stoichiometric model can be collapsed into a system of solvable linear equations. A properly setup stoichiometric matrix is therefore a mathematical representation of the flow of metabolites through a metabolic network and plays a crucial role in developing models based on metabolic knowledge [23].

Of note is that the number of metabolites m will be far lower then the number of reactions n, giving the matrix a vast degree of freedom, and making the equation set underdetermined and unsolvable in its current state. Therefore more constraints are defined mathematically and implemented in order to restrict the equation set and allow for possible solutions [24].

2.2.4 Mathematical implementation of constraints

Mathematical implementation of constraints take the form of either an equation or an inequality. In order to apply the proper mathematical formulation on a constraint it has to be either defined as a ***bound*** or as an ***equality***.

An equality is represented as an equation. The classical example being conservation of mass. According to the *Steady State Assumption* there can be no accumulation or depletion of cellular components and thus all production and depletion of individual components must be equal. This assumption is mathematically represented as the earlier mentioned equation

$$\mathbf{Sv} = \mathbf{0},$$

mentioned in Axiom 5, where S is the stoichiometric matrix and $\mathbf{v}$ is the vector of reaction rate [22]. This is further explored in chapter 2.2.3.

Bounds are represented as inequalities. These constraints limit certain parameters or individual variables, such as concentration, fluxes, or kinetic constants. Limits can be applied to individual fluxes to keep them from leaving their defined possible range in the form

$$v_{min} \leq v \leq v_{max}.$$

Irreversible reactions can be mathematically defined by setting the lower limit of the reaction to 0

$$v_{min} = 0.$$

Concentration of biomass components can never go in the negatives

$$0 \leq x_i.$$

Certain components can have an upper limit due to solvent capacity constraints

$$x_i \leq x_{max}.$$

Kinetic constants can never go in the negative and have an upper limit defined by collision frequency

$$0 \leq k \leq k_{max} \text{ [22]}.$$

Once these constraints are properly mathematically defined, they can be applied to the network reconstruction in order to limit the amount of possible functional states that the network can reach. This is performed in a successive fashion, where the starting point can be represented as a space where the axis are the fluxes through reactions in the network. First the steady state assumption is applied, creating a subspace shown as a hyperplane. The second step is to define the reaction directions by accounting for irreversible reactions and defining each flux as positive. This converts the plane from the steady-state assumption into a semi-finite convex cone shape, the edges of which represent the extreme states. The third step is to apply the capacity constraints to the reaction fluxes. This closes the cone and creates a closed solution space, as a bounded convex subset, in which all the allowable network states lie. After this, by applying more constraints and limitations, the solution space can be further shrunk down to increase the predictive accuracy of the optimization [22]. Fig. 2.3 is a visual representation of this process. After the solution space has been defined, optimization can applied in order to find particular solutions within this solution space.

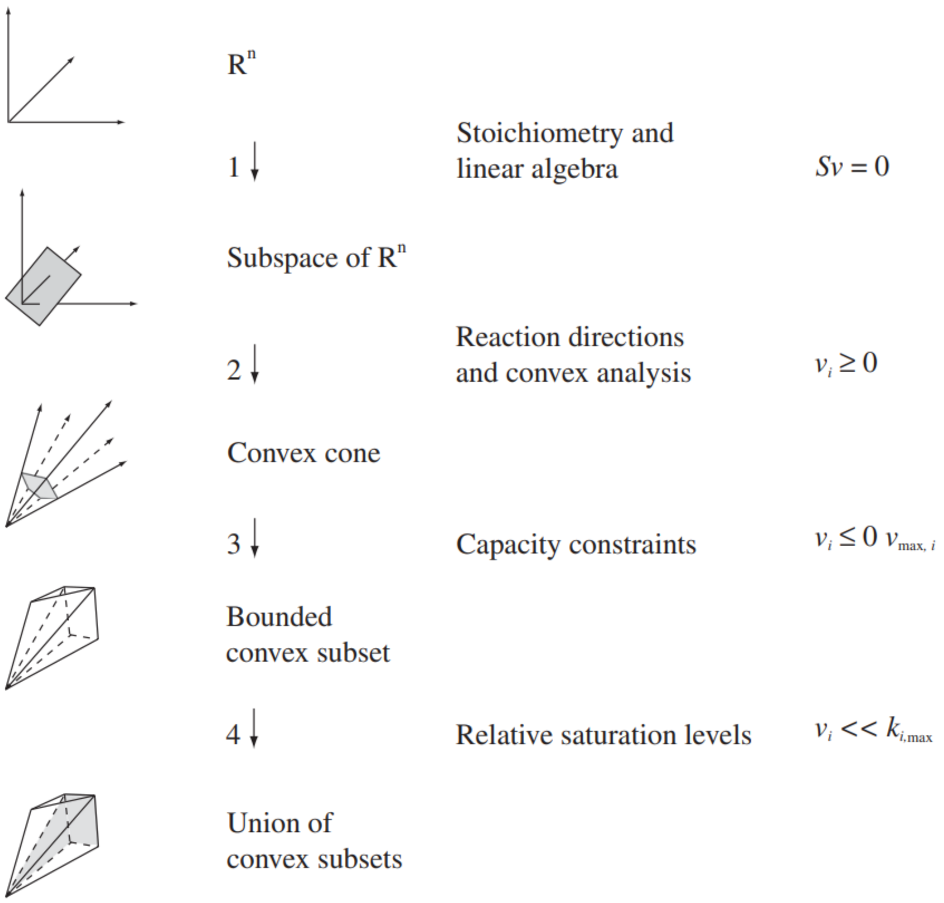


Figure 2.3: A visual representation of the shrinking solution space caused by applying the proper constraints and limitations to the metabolic network outlined previously. It starts in a 3 dimensional space which then has a subspace created in it by applying the stoichiometric constraints and linear algebra to it. After the subspace has been defined, it is further restricted into a cone by adding in the reaction direction constraints and performing a convex analysis. This then forms a solution space represented as an infinite convex cone. The cone is then bounded by applying the capacity constraints and the upper limits of the reactions. From this a bounded solution space has been formed where particular solutions can be represented. By applying further restrictions the space can be bounded and restricted further, allowing for elimination of even more unfeasible network states. Figure taken from Ref. [25].

2.2.5 Linear Programming

Linear programming is the branch of applied mathematics for solving optimization problems of a particular form. It can be defined as the minimisation or maximisation of an objective function subject to a number of inequality constraints. The general form of a linear programming problem is written as

$$\text{maximise} \sum_{j=1}^{n} c_j x_j, \tag{2.4}$$

$$\text{s.t.} \sum_{j=1}^{n} a_{ij} x_j \leq b_i, \;\; i = 1, 2, ...m \tag{2.5}$$

$$x_j \geq 0, \;\; j = 1, 2, ...n \tag{2.6}$$

m are the variables of the problem, while n defines the constraints. c_j represents the objective function coefficient of x_j which represents the decision variables of the objective function. a_i and b_i refers to problem data that defines the problem. A linear programming problem can also be written in matrix form, in which case it uses the general form

$$\text{maximise } z = \mathbf{c}^T \mathbf{x} \tag{2.7}$$

$$\text{s.t. } \mathbf{Ax} \leq \mathbf{b} \tag{2.8}$$

$$\mathbf{x} \geq 0. \tag{2.9}$$

In this form

$$\mathbf{c} = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix}, \mathbf{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}, \mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \tag{2.10}$$

are column vectors that define the constraints and objective function. $\mathbf{c}^T$ is the transpose of vector $\mathbf{c}^T$ and $\mathbf{A}$ is the $m \times n$ matrix where each element is defined as $a_{i,j}$.

Any vector x that satisfies the constraints of the problem is said to be a feasible solution and from this linear problems can be classified into three different classes. The first class are infeasible problems, where there exists no vector x that satisfies the constraints. The second class are unbounded problems, where the constraints do not sufficiently constrain the objective function. This means that there will always be a better possible solution and the solution space stretches infinitely. The third class are bounded problems, where the problem is neither unbounded nor infeasible. This means that there will be an unique solution to the objective function. Note that this does not mean that the variables that yield the optimal objective function are also unique [26].

From subsection 2.2.4 it can be seen that the linear optimization problem that GEM produces is a bounded linear problem. A method for solving these problems have been developed called the *simplex method* .

The simplex method is based on the idea that the optimal solution can be found for a linear programming problem by defining an initial solution and then iteratively improving it. This is achieved by identifying an entering variable that can be increased to improve the objective function and a leaving variable that can be decreased to maintain the feasibility of the solution. This process repeats until an optimal solution is found [27]. A graphical representation of this process can be seen in Fig. 2.4.

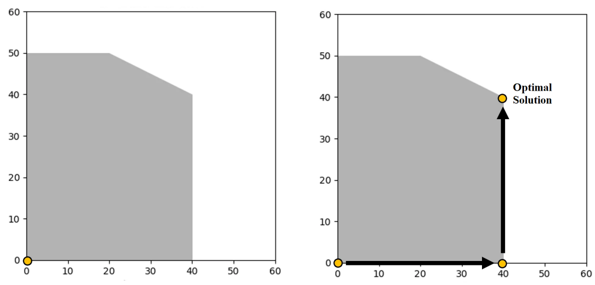


Figure 2.4: Graphical illustration of how the Simplex Method iteratively finds the optimal solution from another feasible solution. It firsts selects 0, 0 as a feasible solution before iteratively improving it along the x-axis, changing to the y-axis once the solution reaches the edge. It keeps iterating until it can no longer improve the solution. This is a general example and does not use any specific data. Figure modified from Ref. [28].

2.2.6 Flux Balance Analysis

As mentioned in 2.2.3, the equation set formed from the stoichiometric model and the steady state assumption is severely underdetermined owing to the many more reactions then metabolites in the matrix [29]. While the constraints implemented in 2.2.4 creates a bounded solution space for the model, a method still remains for arriving at specific solutions. Many methods have been tried, however FBA has risen as the principal method for predicting flux distributions and so find specific solutions in the solution space.

FBA can be applied to predict solutions in this space in the form of calculated flux distributions. By applying the principals of linear programming and the mathematical constraints outlined in 2.2.4, the general vector structure of FBA can be given as

$$\max Z = \mathbf{c}^T \mathbf{v}, \tag{2.11}$$

$$\text{s. t. } \mathbf{S}\mathbf{v} = 0, \tag{2.12}$$

$$lb \leq \mathbf{v} \leq ub. \tag{2.13}$$

In these equations, $\mathbf{c}$ is the vector of weights, or the coefficient vector, that indicates how much the reactions, given as the vector $\mathbf{v}$, contribute to the objective. Eq. 2.12 is the steady state assumption and helps define the system as a set of linear equation. Eq. 2.13 defines the solution space [30].

In order for FBA to produce solutions in the solution space an objective function has to also be defined. The general form of this objective function can be written as

$$Z = \langle \mathbf{w}, \left(\begin{array}{c} \mathbf{v} \\ \mathbf{b} \end{array} \right) \rangle = \sum_i w_i v_i + \sum_j w_j b_j, \tag{2.14}$$

where $\mathbf{w}$ is the vector of weights (w_i) on the internal (v_i) and exchange fluxes (b_i) respectively. The weights define the importance of specific aspects of the objective function [31]. A visual example of this solution space with a particular optimal can be seen in Fig. 2.5.

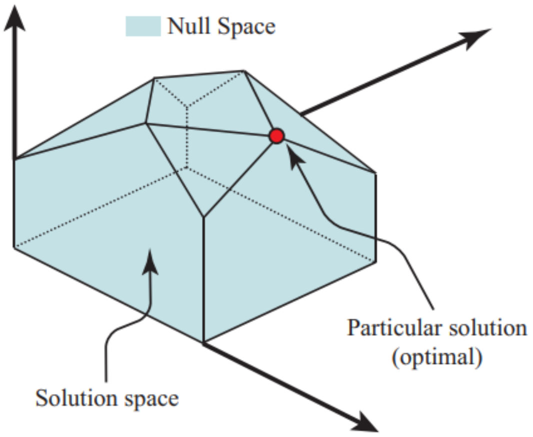


Figure 2.5: Illustration of the bounded null space, and a particular solution found by linear optimization and programming. This figure shows that the optimal solution is found at the edges of the solution space. Not that this is a general example and not a specific one. Figure taken from Ref. [25].

This gives a general structure for finding maximising or minimising an objective function, however, a specific one still needs to be defined. By selecting one that is evolutionary motivated, FBA is more likely to give accurate flux distribution.

For the majority of FBAs this objective is set as the biomass objective function, a representation of the cell producing biomass and growing. It is a reasonable assumption that evolution has optimized the metabolic system of the organism around biomass production as it is a driving factor for replication and life [32]. As such we can assume that the cells metabolism will be under selection pressure to favor internal pathways that optimize production or efficiency of production of biomass and will favor a metabolic network that maximises this. The biomass objective function and its implications is explored further in 2.2.9.

The calculated flux distribution produced by FBA can be used to interpret the metabolic capabilities of a cell under different conditions and modifications [33]. The fluxes act as a representation of the netto flow of metabolites through the reaction in question. However, in order for FBA to give an assumed optimal state of the metabolic network, some assumptions have to be made. One of these is the earlier mentioned assumed steady state [34]. This means that there is no overall accumulation or depletion of metabolites inside the metabolic network, with the exception of the biomass objective function or for exchange reactions. This is a reasonable assumption over a large time scale however over shorter time scales cells can violate this assumption in order to achieve short term goals. Another assumption made for FBA is that the metabolic system has reached its optimal state, focused on the reaction set as the objective function.

As mentioned earlier, the equation set that FBA gives a solution for is underdetermined. While the additional constraints, as well as the selection of a specific objective function, helps to make the equation set solvable, FBA is still likely to find multiple different optimal solutions [35]. Additionally, the limited knowledge on exact flux constraints as well as the large degree of interconnectivity the metabolic network displays will further increase the possible optimal solutions [36]. A graphical visualisation is that the objective function will be parallel to a plane of possible different constraints. In order to achieve more metabolically realistic results for the internal metabolic fluxes, alternative methods and expansions for FBA have been developed.

2.2.7 Parsimonious Flux Balance Analysis

Parsimonious flux balance analysis (pFBA) is an expansion of the FBA method, that seeks to further restrict the number of possible solutions that FBA finds. The underlying assumption for pFBA is that growth pressure will cause a selection for strains that process the growth substrates the most rapidly and efficiently with the minimum amount of used enzyme. We can define this statement as an assumption that cells will be selectively pressured for the fastest growing strains and that the sum of fluxes throughout the network can act as a proxy for minimization of enzyme usage. To approximate this assumption into a functional method, pFBA first employs FBA to optimize growth rate, followed by minimizing the net metabolic flux through the gene-associated reactions in the network. Defined mathematically this step is represented as

$$\text{min} \sum_{j=1}^{m} v_{irrev,\, j}, \tag{2.15}$$

$$\text{s.t. max } v_{biomass} = v_{biomass,\, lb}, \tag{2.16}$$

$$\text{s.t. } \mathbf{S}_{irrev} * v_{irrev} = 0, \tag{2.17}$$

$$0 \leq *v_{irrev,\, j} \leq v_{max}, \tag{2.18}$$

where m are the gene-associated reversible reactions in the metabolic network, max $v_{biomass}$ is the optimal growth rate, found through FBA, and the lower bound of the growth rate is set to this optimal to restrict the solution [37].

By employing pFBA to find the most efficient metabolic network topology, one can find an optimal solution for the objective function that at the same displays the most efficient metabolic network topology [38]. This is likely to give a more accurate solution as the standard FBA method naively finds the most optimal solution for the objective function, but ignores the other aspects of the cell. Other methods have been developed in order to get more accurate predictions then FBA can offer, usually by accounting for a particular situation in which one of FBAs assumptions does not apply.

2.2.8 Minimization of Metabolic Adjustment

One earlier mentioned assumption of FBA is that it assumes the metabolic system is in its optimal state for the objective function set [39]. While this can be a reasonable assumption over a large time scale, for shorter time scales, where the organism have not been exposed to long-term evolutionary pressure, this assumption is often inaccurate. As such the Minimization of Metabolic Adjustment (MOMA) method was developed, which is a method for finding the minimal distance from a point in the feasible space to another given optimum, often outside of the feasible space. The main use case of MOMA was for creating a more accurate calculated flux distribution following a genetic perturbation, such as gene knockouts. The main hypothesis is that the mutant network, following a perturbation, is minimally different from the unperturbed network.

In order to find this minimal distance mathematically MOMA relies on the concept of Euclidean distance. It aims to minimize this distance in the solution space, following the mutant knockout, between a feasible point and the optimal solution of the wild type. The Euclidean distance is defined mathematically as

$$D(w, x) = \sqrt{\sum_{i=1}^{N} (w_i - x_i)^2}. \tag{2.19}$$

Here, D represents the distance in the Euclidean space, N is the space of fluxes, w represents the flux vector of the wild type and x represents the flux vector that is the target of minimization. As this is a non-linear problem, linear optimization will not apply. Thus MOMA relies on quadratic programming (QP) to solve the problem, which uses linear constraints, but can be applied to a quadratic objective function. A general structure for a QP problem is defined as

$$f(x) = \mathbf{Lx} + \frac{1}{2}\mathbf{x}^T\mathbf{Qx}, \tag{2.20}$$

where f(x) represents the objective function, $\mathbf{x}^T$ is the transpose of $\mathbf{x}$, $\mathbf{L}$ is the vector defined as $n \times 1$ and $\mathbf{Q}$ is the matrix $n \times n$. $\mathbf{L}$ represents the linear part of the objective function while $\mathbf{Q}$ is a representation of the quadratic part. In order to transform the minimization of Euclidean distance into a QP objective function, one can observe that minimizing D is equivalent to minimizing its square. Furthermore, $\mathbf{Q}$ can be set to the identity matrix of $\mathbf{I}$ and $\mathbf{L}$ can be set to -$\mathbf{v}$. This, in turn, means we can express MOMA as a quadratic problem by setting

$$min\; f(x) = \frac{1}{2}\mathbf{xIx}^T + (-\mathbf{v})\mathbf{x}, \tag{2.21}$$

$$s.t. \;\; \mathbf{Sx} = 0, \tag{2.22}$$

$$v_i^{'} = 0, \tag{2.23}$$

$$\mathbf{lb} \leq \mathbf{b} \leq \mathbf{ub}, \tag{2.24}$$

where $v_i^{'}$ represents the flux through the knock out of reaction j.

One observation to make is that MOMA finds the point in the solution space that is the closest to the wild-type point, as calculated with FBA. From this we can intuit that, if an FBA solution exists for the model, then a MOMA solution will also exist [40].

While MOMA is intended, and designed for, predictions of the metabolic network after a genomic perturbation, it can also be applied to find the closest feasible solution for a model to experimental data. This is due to the nature of MOMA where it finds the closest point in the solution space of a model to another given reference point. By setting the reference point of the MOMA problem to the experimental data, MOMA can provide a feasible solution that is closest in Euclidean distance to this data.

MOMA is based on the same stochiometric constraints as FBA, but relaxes the assumption of optimal growth. That way it is capable of finding suboptimal flux distributions, that are an intermediate between the models own optimum and the wild type optimum. The objective function of MOMA is therefore not exclusively based on biomass growth like it is for FBA.

2.2.9 Biomass Objective Function

As explained earlier, for most GEMs, cell growth is the target for optimization. Growth can be represented by experimentally determining the metabolites, and their amounts, necessary to

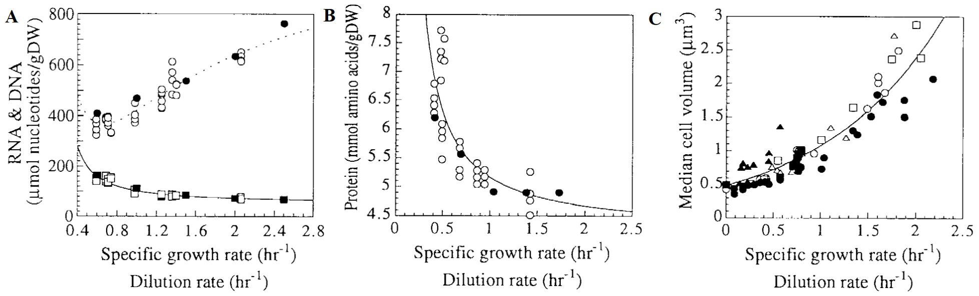


Figure 2.6: Graphical representation of the variance of macro-molecular composition, for RNA, DNA and proteins, as well as cell volume with regards to growth rate. A: The variance for RNA, in the dotted line, and for DNA, with the whole line, with specific growth rate. The circles represent the RNA values and the squares DNA values. The difference in color between them represents where the data is collected from, as Pramanik and Keasling used two different sources of data for these values. From this we can observe that the macro-molecular fraction of RNA increases with specific growth rate, while the fraction of DNA decreases. B: The variance for the molecular amount of amino acids in the biomass with growth rate. The difference in color between the dots represent the different sources of data used for these plots. From these values we can see that macro-molecular fraction of proteins decreases sharply with growth rate. C: The variance of the cell volume with specific growth rate. The different symbols represent the different sources for the data used to produce these graphs. From this we can see that the volume of the cell displays a non-linear increase with the specific growth rate of the cell. This figure was modified from Ref. [12]

synthesize the biomass in the organism. These metabolites can then be represented as a set of reaction fluxes that drain the appropriate metabolites with the correct amounts. The requirements are usually based on the measured values of the biomass composition. The sum of these requirements creates the biomass objective function, which is added in as an extra column in the stoichiometric matrix of the model. This function represents the organism using its metabolites and pathways to produce mass and be capable of life and replicating itself. Due to the nature of these compositions the biomass objective functions are unique to their respective cells [41].

The composition of an *E. coli*s biomass, can be split up into six different groups. These groups are proteins, RNA, DNA, lipids, co-factors and metallic ions, which can be further divided up into subunits for each group [42].

Of note is that a general assumption is made for GEMs and FBA that the biomass composition, which can be represented as the fractions of these groups, is constant and does not change for the organism in question. This is false as we know that the biomass composition is affected by different factors like growth rate and environment. The impact of growth rate on the groups in the biomass composition has been studied closely by Pramanik and Keasling.

They found that the biomass composition shows variance and is affected by the growth rate [12]. These variations for the macro-molecular groups of RNA, DNA and proteins as well as for the volume of the cell can be seen in Fig. 2.6.

Another factor adding uncertainty to the composition of the biomass is the natural limitations caused by the measurement techniques, the temporal and varying nature of the metabolism and the heterogeneous nature of cell populations. A previous effort into studying these variations found the biomass coefficients, and thus the composition of the components in the biomass, to be binomally distributed. To study how this variance would affect the model a study was performed by Maranas et al., studying the impact on the variance on a standard COBRA *E.coli* model. They found that, while the predicted biomass growth was barely affected, the predicted internal flux distribution of the *E.coli* showed major variance as a consequence of the composition variance[13].

In order to make the biomass objective function more accurate, and thereby increase the accuracy of the internal flux distribution, these variations based on can be implemented into the biomass objective function. This can be done by reconstructing the biomass objective function based on the equations outlined.

The biomass objective function is balanced around representing the biomass at $1\ \mathrm{g\ mmol^{-1}}$, as it is required in order to convert the flux prediction to growth rate This is due to the growth rate being given as $\mathrm{g\ gDW^{-1}h^{-1}}$ while the fluxes are given as $\mathrm{mmol\ gDW^{-1}h^{-1}}$. As such any change in the biomass objective function has to account for this by scaling the coefficients of the biomass reaction until the appropriate molar weight of biomass is reached. This is due to a small change in the molar weight of biomass can cause massive perturbations upstream in the flux distribution of the metabolic network and so make the predicted flux values inaccurate.

2.2.10 iML1515 model

The iML1515 model is a model based on the metabolic network of *E. coli* K-12 MG1655 [43]. The biomass objective function is based on data collected by Neidhardt and etc., where they found that the average *E. coli* cell growing at 37 °C in a glucose minimum media with a doubling time of 40 minutes (1.67 μ) has a dry weight of $2.8 * 10^{-13}$g. This dry weight is split into the metabolic groups of protein, at 55% of the dry weight, RNA, at 20.5% of the dry weight, DNA, at 3.1% of the dry weight, and lipid, at 9.1 % of the dry weight. In addition to these there are also the metabolic groups of lipopolysaccharides, peptidoglycan, glycogen, polyamines, cofactors and metallic ions [44]. However, these groups show little to no change with growth rate, as shown in 2.2.9. This is the general biomass composition that is represented in the iML1515 model, through the implementation of the BOF.

2.2.11 Enzyme-constrained models

One of the methods to enhance the standard GEMs is to combine these with the Enzymatic Constraints and Kinetics and Omics data (GECKO) method, in order to further restrict the solution space. It does this by incorporating enzymatic and kinetic data in the model as an additional set of constraints. Enzymes are catalysts in the organism controlling and influencing the metabolic reactions occurring within said organisms. Reactions have a basic constraint in that they can not exceed the reactions maximum rate (V_{max}) defined by the enzymes turnover rate (k_{cat}) multiplied by the enzymes intracellular concentration.

This is represented in the equation

$$v_j \leq k_{cat}^{ij} * [\mathrm{E}_i], \tag{2.25}$$

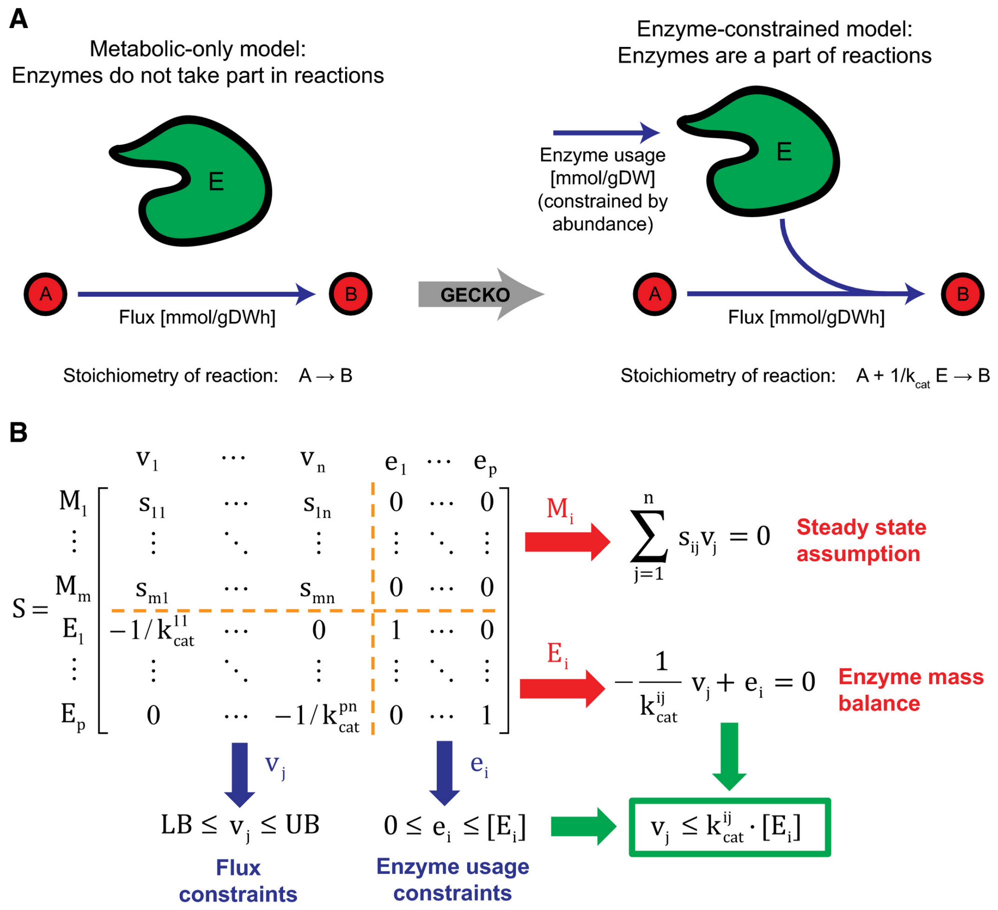
where i represents the metabolite and j represents a corresponding reaction.

Enzymes also have their own mass balance constraint, which can be defined by the equation

$$-\frac{1}{k_{cat}^{ij}} v_j + e_i = 0. \tag{2.26}$$

v_j represents the flux of reaction j, k_{cat}^{ij} is the turnover number for enzyme i catalyzing reaction j. e_i represents the usage of enzyme i.

In order to expand the standard GEM model and account for these enzymatic constraints, the GECKO model expands the stoichiometric matrix. It adds more rows that represents the different enzymes in the system. The turnover rates for each enzyme is included, in order to transform the metabolic flux into enzyme usage. As such the enzymes can be considered ”faux” metabolites that are also consumed for the different reactions. Additionally, more columns are added to the matrix which contains the enzymatic usage value for each enzyme. After the expansion of the stoichiometric matrix is finished, the constraints outlined earlier are then applied. This shrinks the solution space further and the enzymatic data has been successfully integrated into the model. Of note is that the turnover rates, the k_{cat} values, are manually curated and tuned after a draft model has been made. This is to ensure that the model is capable of representing the metabolic phenotypes and to make the model fit with experimental data appropriately [45]. Fig. 2.7 is an illustration of the mathematical process of creating an enzyme constrained model by expanding the stoichiometric matrix to include proteomic data.


A
Metabolic-only model:
Enzymes do not take part in reactions
E
A
B
Flux [mmol/gDWh]
Stoichiometry of reaction: A → B
GECKO
Enzyme-constrained model:
Enzymes are a part of reactions
Enzyme usage [mmol/gDW] (constrained by abundance)
E
A
B
Flux [mmol/gDWh]
Stoichiometry of reaction: A + 1/k_cat E → B
B
S =
M_i
Steady state assumption
E_i
Enzyme mass balance
v_j
e_i
LB ≤ v_j ≤ UB
Flux constraints
0 ≤ e_i ≤ [E_i]
Enzyme usage constraints
v_j ≤ k_cat^ij · [E_i]

$$
S = \begin{array}{c|cccc|ccc}
 & v_1 & \cdots & v_n & & e_1 & \cdots & e_p \\
\hline
M_1 & s_{11} & \cdots & s_{1n} & & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & & \vdots & \ddots & \vdots \\
M_m & s_{m1} & \cdots & s_{mn} & & 0 & \cdots & 0 \\
E_1 & -1/k_{cat}^{11} & \cdots & 0 & & 1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & & \vdots & \ddots & \vdots \\
E_p & 0 & \cdots & -1/k_{cat}^{pn} & & 0 & \cdots & 1
\end{array}
$$

$$\sum_{j=1}^{n} s_{ij} v_j = 0$$

$$-\frac{1}{k_{cat}^{ij}} v_j + e_i = 0$$

$$LB \le v_j \le UB \qquad 0 \le e_i \le [E_i] \qquad \boxed{v_j \le k_{cat}^{ij} \cdot [E_i]}$$

Figure 2.7: A: Illustration of how the basic flux of a reaction is altered in order to implement the enzyme usage as an additional method of constraining the model. B: A representation of how the basic stoichiometric matrix (upper-left sub-matrix) is expanded to include the enzymatic data. S represents the expanded stoichiometric matrix, M refers to the metabolites, E to the enzymes, e to the enzymatic usage and v to the flux through a reaction. Note that the upper right sub-matrix will be empty as the metabolites have all zero enzymatic usage. Figure taken from Ref. [45].

Chapter 3

Software and methods

This chapter provides a summary, along with a short description, of the different softwares and methods used throughout this project. It begins with a description of the various Python packages and software applications used and an explanation on their role in this project. Afterwards, it provides an explanation of how the propagation of uncertainty analysis was adapted to work with the eciML1515 model and the process of performing said analysis on that model. Finally it explores the method of establishing and implementing growth rate-specific BOFs for the iML1515 and the statistical methods utilised to compare it to the experimental data.

3.1 Python packages and software used

Python is an interpreted, object-oriented programming language, with a readability focused syntax [46]. It was used as the programming language for this project with Jupyter Notebook as the interface. The program and language was provided by the Anaconda package and environment system [47]. The following packages and software was used in conjunction with Python for altering the bounds of the models, creating predicted flux distributions, sampling and extracting coefficient data from the models.

3.1.1 COBRAPy

COBRAPy is a Python package used for building and analyzing GEMs. It was used regularly throughout this project to interface with, modify and create predicted flux distributions for the genome-scale and enzyme-constrained models [48].

3.1.2 ReFramed

ReFramed is a Python package that is similarly used for simulating GEMs [49]. It was used in in this project for its MOMA function, as the expanded functionality of it was better for the purpose of this project then the one provided by COBRAPy.

3.1.3 Gurobi

Gurobi is a proprietary solver designed for mathematical optimization. It was used to solve the different optimization problems occurring throughout this project, such as the linear programming problems for FBA and pFBA as well as the quadratic programming problems for minimzation of metabolic adjustment (MOMA) [50].

3.2 Quantifying the propagation of uncertainty from biomass composition to flux predictions

This section will outline the steps necessary to perform a uncertainty propagation analysis on the eciML1515 model, an enzyme constrained variant of the iML1515 model, which is a comprehensive metabolic COBRA model of an *E.coli* cell. The works of Maranas et al., where they performed a similar analysis on the iML1515 model, was the inspiration for the methods utilised in this section [13].

3.2.1 Standardizing the eciML1515 to the iML1515 for propagation of uncertainty analysis

I used the latest enzyme constrained model for *E. coli*, the eciML1515 [51], for the uncertainty propogation analysis. It was modified to make it more standardised and comparable to the commonly used metabolic model for *E. coli*, the iML1515. The biomass function and objective was replaced with the biomass function from iML1515, as the pathways leading into the biomass function in eciML1515 are categorized based on the enzymes catalyzing the reatctions. This causes some issues with the provided functions for propagation of uncertainty. The reverse beta-oxidation pathway was disabled by setting its upper bound to 0 $\mathrm{mmol\,gDW^{-1}h^{-1}}$, as this pathway is an engineered one that is absent in wild-type strains, but included in the model [52]. The reverse pyruvate synthase pathway was disabled by setting the upper bound of the reaction to 0 $\mathrm{mmol\,gDW^{-1}h^{-1}}$, as this pathway is only induced in superoxide generators [53]. The scripts for adjusting the eciML1515 can be found in the provided Github repository in the folder scripts/ModelAdjustments.

3.2.2 Propogation of uncertainty analysis on the eciML1515 for biomass yield and specific metabolic fluxes

The methods discussed in this section were largely based on, and followed the same procedures as the perturbation analysis performed by Maranas et al. on the iML1515 [13].

After the eciML1515 was modified to cooperate with the scripts provided in the study by Maranas et al. [13] for the perturbation analysis, the parameters for analysis were set up. The parameters subjected to perturbations and uncertainty were the biomass precursor coefficients (Biomass uncertainty), the macromolecular composition of the biomass (Macro uncertainty), the growth associated maintenance of the model (GAM uncertainty) and the non-growth associated maintenance of the model (NGAM uncertainty). In addition certain selected biomass precursors were picked as parameters for individual sampling of a single biomass coefficient. The selected biomass precursors were all the DNA, RNA and the amino acids. The parameters were sampled 10 000 times based on a normal distribution with the mean value corresponding to the original value. This was done four times with a growing relative standard deviation (SD) of 5%, 10%, 20% and 30%. The biomass precursors coefficients and the GAM value were extracted from the biomass reaction with the lumped coefficients, then sampling was performed in accordance with the earlier mentioned parameters.

To normalize the biomass molecular weight (MW) to 1 g mmol^{-1}, following the uncertainty injection of the biomass constituents, the equation

$$\mathbf{c_i^{new}} = \frac{\mathbf{c_i^{samp}}}{\sum_{\mathbf{i\epsilon I_{Biom}}} \mathbf{c_i^{samp}} \mathrm{MW}_\mathbf{i}}, \tag{3.1}$$

was utilised. I_{biom} is the set of biomass precursors used in the biomass reaction, c_i^{samp} is the randomly sampled coefficient for the biomass precursor i, MW_i is the MW of the biomass precursor i and c_i^{new} is the coefficient for the biomass precursor i, calibrated to make the biomass functions MW be equal to 1 $\mathrm{g\ mmol^{-1}}$.

For the Macro uncertainty, the weight fractions of the macromolecular classes were sampled and the coefficients of constituents in each class were uniformly scaled. As for the sampling of the indivdual biomass precursors, in order to ensure a biomass MW at 1 $\mathrm{g\,mmol^{-1}}$, the parametric uncertainty injected into the singular biomass precursor was equally distributed after biomass MW normalization to ensure the relative fraction of precursors was kept. This process was performed using a script provided by Maranas et al., with some modifications in order to make them functional with the eciML1515 model. The modified scripts can be found in the provided Github repository under scripts/Part1.

After the sampling, a pFBA is performed for each sample in order to find a predicted flux distribution. From the pFBA results a mean and standard deviation was calculated for the flux of every reaction. In order to find the propogation of uncertainty throughout the flux distribution a standard deviation ratio (SDR) was calculated for each flux. The calculation was done using the formula

$$SDR_{Input}^{Output} = SDR_c^v = \frac{\Delta(\frac{SD}{Mean})_{Output}}{\Delta(\frac{SD}{Mean})_{Input}}, c \tag{3.2}$$

where v represents the reaction in question, c represents the parameter the uncertainty was applied to and SDR_c^v represents the ratio between the output relative SD with the input relative SD.

To determine the SDR of the total cofactor production fluxes for NADH, NADPH and ATP, the sum of all the fluxes producing these cofactors is calculated. This sum was then set as a new flux for SDR calculation. This was done by extracting the stochiometric values and the flux through the cofactor reactions from the sampling results. The formula to calculate the new flux was

$$v_{cofactor} = \sum_{j\epsilon J_{cofactor}} S_{Cofactor,j}\, v_j. \tag{3.3}$$

$v_{cofactor}$ is the cofactor production flux, $S_{Cofactor,j}$ is the stochiometric value of the cofactor in the reaction j and v_j is the flux through the reaction j. This formula was applied to all the producing reactions for ATP, NADH and NADPH in order to find the netto flux production and SDR of these cofactors.

As the code incurs a heavy workload, they were run using an external computer. The computer was provided by the Network Systems Biology group at NTNU, it had a 8-core processor with two threads per core and a clock speed of 5360 MHz. In addition, while the source code for uncertainty injection and pFBA was provided for all parameters, the scripts for SDR calculation and the ATP, NADH and NADPH cofactor production fluxes were only provided for the injection of uncertainty into all biomass coefficents parameter. Due to this those scrips had to be modified from the script provided for the parameter of all biomass coefficents. Once every script was set up and modified they were run on the eciML1515 model.

After the SDR was calculated for each flux and for each set parameter the data was collected, the appropriate SDR values were extracted and analyzed. They were then compared with the results from the Maranas et al. projects. The full data sets can be found in the provided Github repository.

3.3 Implementing a growth rate-specific biomass objective function for the iML1515 model

This section of the project covers the methods and formulas needed in order to create a condition-dependent and growth rate-specific version of the iML1515 based on provided experimental data from Gerosa et al. study where they collected data on the metabolic flow of *E. coli* using different carbon sources [15].

3.3.1 Establishing the biomass coefficient formulas for a condition-dependent and growth rate-specific biomass composition

The standard iML1515 model was used for the reconstruction. The biomass constituents were extracted from the objective function of the model and then grouped and categorized based on the data provided in supplement (x). These groups are amino-acids (proteins), RNA, DNA, lipids, co-factors and metal ions. To calculate the growth rate-specific coefficient for each group, data from the study by Pramanik and Keasling was utilised, where they established equation for the calculation of a growth rate-specific biomass composition [54]. These formulas are based on experimental data and detailed sampling. As the data provided no info on the impact of growth rate on metal ions and co-factors, they were not considered for the production of a growth rate-specific biomass objective function. The equation for the growth rate-specific RNA for a biomass composition was

$$\mathrm{C}^{\mu}_{\mathrm{RNA}} = 1.1395 - \frac{0.9665}{\mu} * 2^{-0.665/\mu}, \tag{3.4}$$

where C^{μ}_{RNA} is the growth rate-specific molecular fraction of the biomass composition and μ is the specific growth rate. The unit for the RNA molecular fraction is given as $\mathrm{mmol\ NTPs\ gDW^{-1}}$. This equation represents a non-linear increase in the RNA fraction as the specific growth rate rises.

The equation used to calculate a growth rate-specific molecular fraction of the DNA for the biomass composition was

$$\mathrm{C}^{\mu}_{\mathrm{DNA}} = \frac{0.1\mu}{0.023}(2^{0.017+0.663/\mu} - 2^{0.663/\mu}). \tag{3.5}$$

C^{μ}_{DNA} is the growth rate-specific molecular fraction of DNA in the biomass composition and μ is the specific growth rate. The unit for the DNA molecular fraction is given as $\mathrm{mmol\ dNTPs\ gDW^{-1}}$. This equation represents a non-linear decrease in the DNA fraction relative to the increase in the specific growth rate.

The equation used to calculate a growth rate-specific molecular fraction of the proteins for the biomass composition was

$$\mathrm{C}^{\mu}_{\mathrm{Protein}} = 4.228 * (2^{0.288/\mu}). \tag{3.6}$$

$C^{\mu}_{Protein}$ is the growth rate-specific molecular fraction of proteins in the biomass composition and μ is the specific growth rate. The unit for the protein molecular fraction is given as $\mathrm{mmol\ AAs\ gDW^{-1}}$, where AA stands for the amino acids. This equation shows that the growth rate-specific molecular fraction of the proteins in the biomass composition is reverse-proportional with the specific growth rate.

Equations for the growth rate-specific lipid fraction were not provided in the study, although equations for growth rate-specific surface area of an *E. coli* cell were. As the surface area of a cell is directly proportional to the amount of lipids in the cell, due to the structure of the cells membrane wall. The outer membrane of an *E. coli* is an asymmetric bilayer consisting of phospholipids and glycolipids [55]. This can be used to determine a growth rate-specific from experimental data and the equations for the growth rate-specific surface area of the *E. coli* These equations are

$$\mathrm{R} = 0.293 * 2^{(0.41\mu)}, \tag{3.7}$$

which represents the radius, R, of a cell based on a specific growth rate, represented by the μ variable. The length of an *E. coli* based on specific growth rate is given with the equation

$$\mathrm{L} = 2 * 2^{0.333\mu}. \tag{3.8}$$

The length of the cell is given as L, while the specific growth rate is given as μ. By applying the assumption that an *E. coli* cell is a cylindrical shape with hemispherical caps, the surface area can be calculated based on the general equation for the surface area of this structure

$$\mathrm{A} = 2\pi\mathrm{R}(\mathrm{L} - 2\mathrm{R}) + 4\pi\mathrm{R}^2. \tag{3.9}$$

A represents the area, L the length and R the radius of the shape. This gives an equation set for calculating the growth rate-specific surface area of an *E. coli*.

In order to find a growth rate-specific equation for the lipid molecular fraction, experimental data was used. The experimental data contained the specific growth rate of an *E. coli* in addition to the number of lipids in the cell. From this the amount of lipids per surface area was calculated using the equation

$$C^{\mu}_{lipids}\,factor = \frac{n_{lipids}}{A_{cell}}. \tag{3.10}$$

n_{lipids} is the amount of lipid in mmol in the experimental data, A_{cell} is the surface area of the cell in question. The factor between these two was then used to establish a growth rate-specific equation for the amount of lipids in the biomass composition. This equation is

$$C^{new}_{lipids} = 0.2485 * A_{cell}, \tag{3.11}$$

where C^{new}_{lipids} is the new calculated coefficient of the lipid group going into the biomass and A_{cell} is the calculated surface area of the cell based on growth rate.

3.3.2 Implementing and normalizing growth rate-specific stoichiometric coefficients for the construction of a growth rate-specific biomass objective function

Next, a method for converting the coefficients from the standard biomass objective function into growth rate-specific ones was established. The first step to this method was to define a model and specific growth rates as a base. The model to use as a base was the iML1515 and the specific growth rates were taken from the experimental data mentioned earlier. The next step of the method used the above formulas and the specific growth rate to establish growth rate-specific group coefficients for the biomass objective function. Once these are calculated, each of the coefficients for the BOF in the iML1515 were extracted and grouped. The BOF for the iML1515 can be defined as

$$\sum_{i\epsilon I_{Biom}} c_i = I_{Biom}, \tag{3.12}$$

where c_i represents the coefficient of the metabolite i as part of the biomass objective function I_{Biom}. Each of these coefficients were subsequently grouped and categorized based on the metabolite groups of RNA, DNA, lipids and proteins. The sum coefficient of each group was then calculated by using the equation

$$\sum_{i\epsilon I_G} c_i = C_G, \tag{3.13}$$

where C_G denotes the sum of the coefficients of each group G. The next step of the method involved finding the factor between the growth rate-specific group coefficient and the extracted group coefficients. This was done using the equation

$$Q = \frac{C^{\mu}_{(Group)}}{C_{(Group)}}. \tag{3.14}$$

where $C^{\mu}_{(Group)}$ is the calculated growth rate-specific group coefficient and $C_{(Group)}$ is the extracted group coefficient. This factor was then applied uniformly to each individual components coefficient in the group. This scales up or down each coefficient appropriately in order to adjust the sum to the growth rate-specific group coefficients. The formula used to do this was

$$\forall c \, \epsilon \, Group : c^{\mu} = Q * c, \tag{3.15}$$

where c is the individual coefficients of each component in the group and Q is the calculated factor from formula 3.14. This produced new established growth rate-specific groups defined as

$$\sum_{i\epsilon I_G} c^{\mu}_i = C^{\mu}_G. \tag{3.16}$$

c^{μ}_i is the growth rate-specific component coefficient for the metabolite i, C^{μ}_G is the growth rate-specific coefficient for the metabolite group G and I_G is each of the metabolite groups represented in the biomass objective function.

After each of the growth rate-specific coefficients were established, they had to be normalized appropriately in order to make the growth rate-specific BOF represent 1 g mmol^{-1} of the biomass. It was decided that this would be performed by normalizing each of the metabolite groups to 1 g mmol^{-1} . By representing each of the groups as a percentage of the biomass

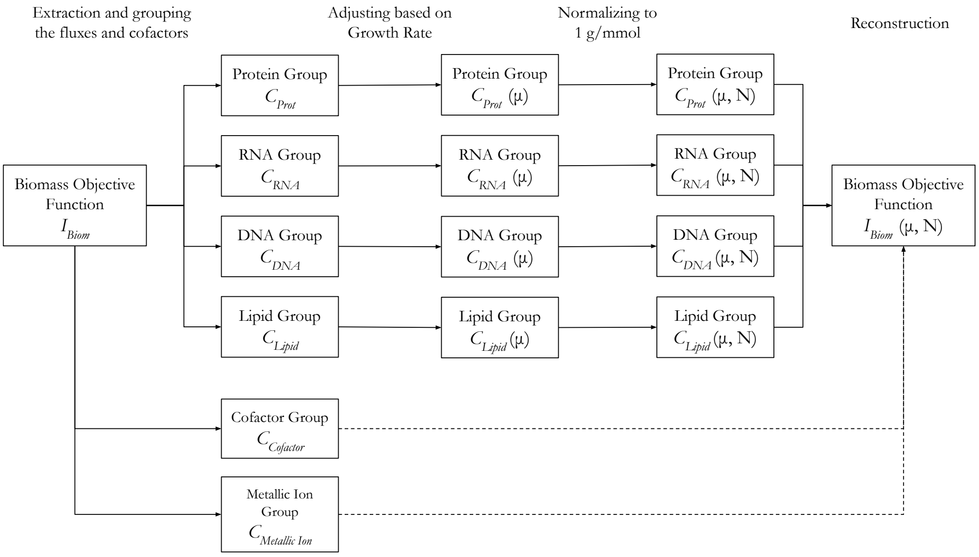

Extraction and grouping the fluxes and cofactors
Adjusting based on Growth Rate
Normalizing to 1 g/mmol
Reconstruction
Biomass Objective Function I_{Biom}
Protein Group C_{Prot}
RNA Group C_{RNA}
DNA Group C_{DNA}
Lipid Group C_{Lipid}
Protein Group C_{Prot} (μ)
RNA Group C_{RNA} (μ)
DNA Group C_{DNA} (μ)
Lipid Group C_{Lipid}(μ)
Protein Group C_{Prot} (μ, N)
RNA Group C_{RNA} (μ, N)
DNA Group C_{DNA} (μ, N)
Lipid Group C_{Lipid}(μ, N)
Cofactor Group $C_{Cofactor}$
Metallic Ion Group $C_{Metallic\ Ion}$
Biomass Objective Function I_{Biom} (μ, N)

Figure 3.1: A visual representation of the workflow for establishing a growth rate-specific biomass objective function and normalizing it to 1 g/mmol. This process was performed on the standard iML1515 model. The first step of the process was the extraction and grouping of the fluxes and cofactors from the biomass objective function. The groups were based on the metabolite groups the biomass consists of. After the coefficients were grouped they were subsequently set to the growth rate-specific ones calculated from the equations outlined in 3.3.1. Once each of the groups were converted to be growth rate-specific they were then normalized to 1 g mmol^{-1}. From these group coefficient a new growth rate-specific BOF was set in the model using the groups as its coefficients.

composition, it is then naturally normalized to $1\ \mathrm{g\,mmol}^{-1}$. A similar method was used to normalize the groups as done earlier by calculating a quotient for the component coefficients and MW of the metabolites. The input coefficients multiplied with the MW was adjusted so that the sum of them was equal to $1\ \mathrm{g\,mmol}^{-1}$. The metabolite groups can then subsequently be defined as

$$\sum_{i\epsilon I_G} c_i^{\mu,\,N} = C_G^{\mu,\,N}. \tag{3.17}$$

$C_G^{\mu,\,N}$ represents the normalized growth rate-specific sum coefficient of each group and $c^{\mu,\,N}$ represents the component coefficients of said groups.

Once the normalized growth rate-specific group coefficients were established, new reactions were created in the model corresponding to each group. From these reactions a new growth rate-specific BOF was created and implemented into the model, with the ID: ”NewBiomassReaction”. This new BOF can be defined with the equation

$$\sum_{G\epsilon G_{Biom}} C_G^{\mu,\,N} = I_{Biom}^{\mu,\,N}, \tag{3.18}$$

where $I_{Biom}^{\mu,\,N}$ is the growth rate-specific BOF. A visualisation of the workflow for this method can be found in Fig. 3.1 and the code to perform this method can be found in the Github repository.

3.4 Implementing eight different growth rate-specific biomass functions for the iML1515 based on experimental data

Eight different models were created using the script outlined in the previous chapter on the iML1515 model. These eight models were based on the experimental data provided by Gerosa et al. [15]. The details of each model is shown in Table 3.1.

Table 3.1: The details of each of the eight growth rate-specific BOF models created for this project. The models were created using the method outlined in the previous section. This table establishes the name of the models, as well as their specific growth rate, what carbon source the data is based on and specific important exchange reactions. The models can be found in the Github repository under models.

Model ID	$\mu\,(h^{-1})$	**C. Source**	**C. Source Uptake** $(\frac{mmol}{gDWh})$	**Acetate Secretion** $(\frac{mmol}{gDWh})$	**Secondary Secretion** $(\frac{mmol}{gDWh})$
ReBMAc	0.29	Acetate	13.58		
ReBMGal	0.18	Galactose	1.97		
ReBMFru	0.49	Fructose	8.33	3.33	
ReBMSucc	0.51	Succinate	15.90	3.32	Fumarate 1.14
ReBMGlu	0.65	Glucose	9.65	6.83	
ReBMPyr	0.39	Pyruvate	26.71	11.91	Lactate 1.16
ReBMGly	0.49	Glycerol	10.14	0.60	
ReBMGlcn	0.59	Gluconate	5.00	5.00	

The given specific growth rate, μ, was used for the construction of a growth rate-specific BOF. For the uptake reactions, the lower bound for glucose was set to 0 $\mathrm{mmol\,gDW^{-1}h^{-1}}$ to disable that as a carbon source while the lower bound of the carbon source was set to -1000 $\mathrm{mmol\,gDW^{-1}h^{-1}}$ to enable that as a carbon source. For the secretion values the upper bound of the exchange reactions in question were set to 1000 $\mathrm{mmol\,gDW^{-1}h^{-1}}$ to enable secretion of the component. The other exchange reactions outside of the ones mentioned were not modified. A similar set of reference models were created from iML1515, with the same bounds, but without a specific growth rate-dependent biomass objective function. The models were then converted to a ReFramed compatible format, using the provided conversion function in ReFramed, and MOMA was performed. The secretion, uptake and growth rate values from the experimental data was used as reference values for each corresponding model. These values were then extracted from the solution MOMA created and inserted back into the model as bounds for the growth rate and exchange reactions. Some slack was introduced into the bounds by rounding the values up for the upper bound and down for the lower bound. The rounding target was set to the nearest hundredth. After this the models were converted back to COBRAPy for the sampling.

Sampling was performed on each model, using COBRAPy, with the number of samples set to 10 000. This produced 10 000 predicted flux distributions, with variations based on the sampling, for each model. Each individual predicted flux distribution was extracted and the fluxes for the set of reactions corresponding to the fluxes in the experimental data were saved. In order to compare the fluxes from the sampling to the experimental data the root square error between each experimental data and the corresponding sampled flux was found. For each sample the mean of these errors were then given as the root mean squared error (RMSE) of that sample. The formula used to calculate RMSE was

$$RMSE_i = \frac{\sum_{j \epsilon J} \sqrt{(c_{j,i} - c_{j,J})^2}}{n_J}, \tag{3.19}$$

where i is the selected sample, $c_{j,i}$ is the flux through reaction j and J is the set of experimental data for each reaction. n_J is the number of reactions in the experimental data.

After the RMSE was calculated for each sample in each model a t-test was performed between the reference models and the experimental models. This was done to test if there was a statistical significant difference between the two populations and in what direction the difference lies.

A second set of statistical testing was performed, this time with a focus on the relative error. To calculate the relative error for each reaction the equation

$$RE_j = |\frac{(c_{j,i} - c_{j,J})}{c_{j,J}}|, \tag{3.20}$$

was used, where RE_j represents the relative error of the reaction j, $c_{j,i}$ represents the value of the reaction j from sample i and $c_{j,J}$ represents the value of the reaction j from the experimental data J. Each sample was given a relative error score by calculating the mean relative error of each reaction in the sample. These scores were then compared between the growth rate-specific biomass objective function variants and the reference iML1515 models by using a t-test analysis.

Chapter 4

Results and discussion

This chapter presents the results of the project, which was divided into two main parts. The first part involved an analysis on the parametric propagation of uncertainty for the eciML1515 model, in comparison to the results found in a similar study by Maranas et al., which was conducted on the iML1515 model [13]. The study evaluated the uncertainty propagation on growth rate, ATP-, NADH- and NADPH production following injection of parametric uncertainty. The parameters for injection was for all biomass coefficients, the macro-molecular fraction of the biomass composition, the GAM, the NGAM and certain individual biomass coefficients.

The second part was a study on how creating a growth rate-specific (BOF) would impact the accuracy of predicting internal metabolic flux distributions. This was done on the iML1515 model and the specific growth rates, as well as the target flux distribution, were based on experimental data gathered by Gerosa et al. [15].

The equations for a growth rate-specific biomass objective function were sourced from a study performed by Pramanik and Keasling [54].

4.1 Propagation of parametric uncertainty on metabolic flux predictions

The following section covers the results obtained from the propagation of uncertainty results for the eciML1515. The methods used for arriving at these results can be found in 3.2.2 and the full data for the analysis can be found in the Github repository.

4.1.1 Growth rate predictions for eciML1515 following metabolic modifications for parametric uncertainty propagation analysis

In order for the scripts and methods given by Maranas et al. for the uncertainty propagation to work with the eciML1515 model, certain specifications and modifications had to be done on the model. These are laid out in chapter 3.2.1.

After the reverse beta-oxidase and reverse pyruvate synthase pathway was disabled in both models, FBA was performed, with glucose as the limiting metabolite, in order to evaluate the impact of the pathway changes on the metabolic network. The results of the FBA on the biomass growth is laid out in Table 4.1. eciIML1515 showed a significant reduction in growth rate with FBA while the standard iML1515 showed no change. This significant reduction for the growth rate in the eciML1515 model is unexpected as Ref. [56] and Ref. [53] shows that these pathways should not be present in wild type (WT) *E. coli* cells.

Considering how a GECKO model is constructed, as laid out in 2.2.11, it is possible that the models enzyme activity was implemented with the assumption that the reverse beta-oxidase pathway was an active pathway in the metabolic network. Therefore, when the k_{cat} values were manually configured and adapted, these pathways were active. This, in turn, made these pathways necessary for the biomass objective function and made them growth-coupled. As a consequence, when these pathways are disabled, the growth rate of the model displays a massive reduction.

This could possibly lead to further consequences for this analysis as it is possible that other pathways in the metabolic network were also coupled with these pathways. Ideally, the eciML1515 would be further modified, following this discovery, to account for this and to retune the k_{cat} values with these pathways disabled. However, this is a time consuming process, which was a limited resource for this project. As such it was decided that it was beyond the scope of the project and that it would move on with the eciML1515 staying as it is following the disabling of the pathways.

Table 4.1: A representation of the growth rate change between the standard iML1515 and the eciML1515 models after disabling the reverse pyruvate synthase and reverse beta-oxidase pathways. eciML1515 shows a considerable reduction in growth rate, while iML1515 shows no change.

Models	μ before disabling pathways (h^{-1})	μ after disabling pathways (h^{-1})
iML1515	0.8770	0.8770
eciML1515	0.5744	0.1939

4.1.2 Propagation of parametric uncertainty on the growth rate predictions of eciML1515

The modified eciML1515 model was ran through a parametric uncertainty analysis with the parameter set to all biomass coefficients. The setup for the uncertainty analysis is outlined in 3.2. All the references in the computer code to specific reaction IDs had to be altered to match with the reaction IDs in the eciML1515. For reversible reactions, the eciML1515 splits these up into different reactions. This had to also be accounted for by implementing them as separate reactions in the code. The modified scripts can be found in the provided Github repository under scripts/Part1. After the analysis was done the datasets were collected and the SDR value for the biomass growth was calculated. For comparison, the results of the similar analysis performed on the iML1515 was acquired from the study by Maranas et al. [13]. The results of the biomass coefficient parametric uncertainty on the biomass production is displayed in Fig. 4.1.

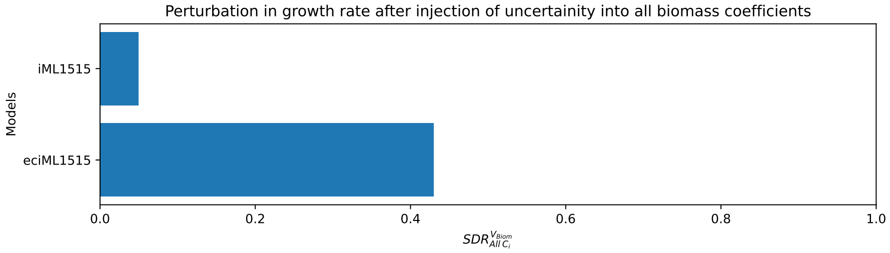


Figure 4.1: Bar plot representing the difference in $SDR_{All\ c_i}^{V_{Biom}}$ between the iML1515 model and eciML1515 model. This is after uncertainty was inserted into the biomass coefficients of both models. Both models displayed an SDR on biomass growth of under 1.0. This means that both models display a dampening effect on the biomass growth after inserting uncertainty into the biomass coefficients of the biomass objective function.

This result that both models have a growth rate SDR, as defined by equation 3.2, of < 1.0. SDR represents the size of the normal distribution on the specified reaction from the sampling compared to the normal distribution on the uncertainty injection on the coefficients. An SDR value of 1.0 means that the reaction displays the same normal distribution as the one applied to the parameter coefficients, while > 1.0 means the reaction displays a wider normal distribution and < 1.0 a tighter one. From this, we can define that an SDR of > 1.0 means that the reaction has an amplifying effect on uncertainty, while an SDR of < 1.0 means the reaction has a dampening effect on uncertainty. Since both models display an $SDR_{All\ c_i}^{V_{Biom}} < 1.0$ this means that the biomass reaction is dampening on the propagation of uncertainty from the biomass coefficients.

From this it can be said that the biomass production of both models is *robust*, meaning it is resistant to perturbations in the biomass coefficients. While both models were expected to display a robust biomass production, owing to the nature of the metabolism and selection pressure, its worth noting the significant difference between them. iML1515 displayed an $SDR_{All\ c_i}^{V_{Biom}}$ of 0.049 while the eciML1515 displayed an $SDR_{All\ c_i}^{V_{Biom}}$ of 0.42. This points to the eciML1515 being far less robust for biomass composition perturbations compared to the iML1515. The reason for this difference could be that the enzyme constrained model has a reduced solution space, compared to the standard GEM model. These enzyme constraints limit the capability of the model to adjust the internal fluxes to counteract perturbations in the biomass composition and maintain the growth rate. This would manifest as a reduced robustness, when looking at the growth rate, for the eciML1515 compared to the iML1515. It is possible that additional constraints would amplify this effect and reduce the solution space of the model further.

Although, as mentioned above, the overall growth rate of the eciML1515 is lower then the iML1515, as a result of the pathway alterations. This could have had an impact on this result owing to the larger impact any absolute variation of the biomass growth would cause in the calculated SDR. This is likely to cause an inflation of the SDR for the eciML1515, although it is difficult to say for certain how big this inflation would be. Additionally, it is possible that the disabling of these pathways also effected other pathways, by them potentially being coupled as a result of the kcat values being tuned with these active. This could further restrict the solution space and inflate the SDR as a result of this. In order to study this the analysis could be performed again on a retuned version of eciML1515 that does not rely on the pathways discussed in chapter 4.1.1, however, as mentioned, that would be an undertaking beyond the reach of this project.

After the parametric uncertainty analysis on all biomass coefficients ($SDR_{All\ c_i}$) was done, a new analysis was performed with the parameters set to macromolecular fractions in the biomass composition (Macro), GAM and NGAM. The results of the perturbation analysis on the biomass growth rate from these parameters can be seen in Fig. 4.2. The Macro uncertainty injection showed a large impact on the SDR of the growth rate and displayed an amplifying effect. This was an unexpected result as the macromolecular fractions should display a similar level of robustness to the biomass coefficients, albeit it is likely to display a slightly higher value for SDR. This is due to standard deviation insertion of the macromolecular fractions prioritising the largest fractions of the biomass. This means that the coefficients that are the most abundant in the biomass gets prioritised and likely leads to a higher SDR on the biomass growth. This can be seen in the study by Maranas et al. where they performed this analysis on the iML1515 [13]. From this study we can see that the iML1515 still displays a dampening effect on the Macro parameter and only has a slightly higher SDR then for the all biomass coefficient parameter. The reason why the results display a massive SDR on eciML1515 for biomass growth with the macro parameter is not clear. One possible explanation is that the code for insertion of uncertainty for the single coefficients-, macro-, GAM- and NGAM parameter had to be modified from the code belonging to the all biomass coefficient parameter. This was a result of the Maranas et al. [13] study not providing enough of the necessary source code in order to replicate their results. Therefore, the computer code to perform this part of the analysis had to be adapted from the earlier code as well as the information they presented in their study. As such it is possible these scripts are not working correctly and giving the wrong values. If so it is likely that the Macro results will display unexpectedly high values for the other SDR results as well.

The GAM and NGAM uncertainty injections, on the other hand, displayed a minimal effect on the SDR of the biomass production, with both displaying a major dampening effect. The NGAM uncertainty was expected to not have much of an impact on the growth rate, owing to the fact that it is tied to maintenance reactions that are not growth associated. As such any variance on the NGAM value should have a minimal impact on the growth rate. This is reflected in both the results for the eciML1515 and the results from Maranas et al. [13] study where the SDR on growth rate relative to uncertainty injection of the NGAM value was 0.02. The GAM uncertainty injection displaying a major dampening effect on the growth rate is, on the other hand, unexpected. The GAM is directly tied to the growth of the cell and, as such, any variance in the GAM value is expected to have a major impact on the growth rate. This result can be seen in the results from Maranas et al. study where they found that the GAM injection of

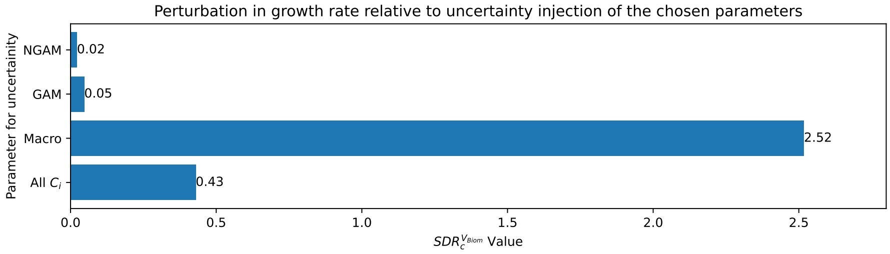


Figure 4.2: Bar plot representing the $SDR_{All\ c_i}^{V_{Biom}}$ after inserting uncertainty into different parameters. The parameters are NGAM, GAM, macromolecular fractions (Macro) and all biomass coefficents (All C^i) for the eciML1515 model. The SDR of the Macro parameter was 2.52, while the SDR for all biomass coefficients, GAM and NGAM was 0.43, 0.05 and 0.02 respectively. This points to the Macro parameter displaying an amplifying effect on propagation of uncertainty, when looking at the growth rate. The other parameters displayed dampening effects.

uncertainty was the parameter that displayed the highest SDR for growth rate. The reason why the eciML1515 does not share this result is unclear. It is possible that the earlier mentioned implementation of the code for performing this injection is faulty, as the source code was not provided.

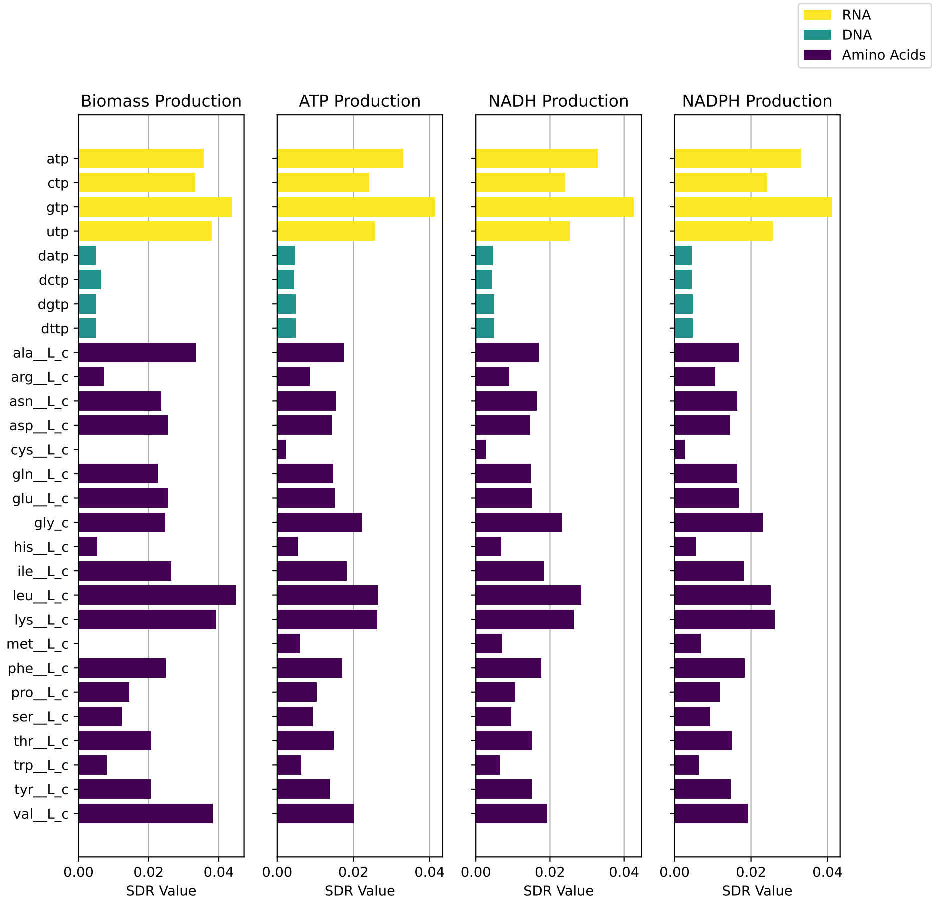


Figure 4.3: Results from the uncertainty injection into specific biomass constituents, looking at the propagation of uncertainty on the production of biomass, ATP, NADH and NADPH. The specific biomass constituents selected were the RNAs, DNAs and amino acids. All the coefficients showed overall little impact on the uncertainty propagation.

4.1.3 Propagation of uncertainty on biomass production, ATP production, NADH production and NADPH production based on injection of uncertainty into specific individual biomass constituents

After the propagation of uncertainty for all biomass coefficients was studied, the study focused on the propagation in a single biomass coefficient at a time. This was done by performing an uncertainty insertion with a single biomass coefficient as parameter, as laid out in 3.2.2. pFBA was performed on each sample and the SDR value of each flux was calculated. The SDR values for growth rate, ATP production, NAD production and NADH production can be seen in Fig. 4.3.

These result show that the SDR impact of individual biomass coefficients on the growth rate is severely dampened, with most values gathered around 0.03 $SDR_{c_i}^{V_{Biom}}$ compared to the injected SDR on the individual biomass coefficient. This is in stark contrast to the 0.43 $SDR_{All\ c_i}^{V_{Biom}}$ from the previous injection result. This contrast is expected as, a single biomass coefficient

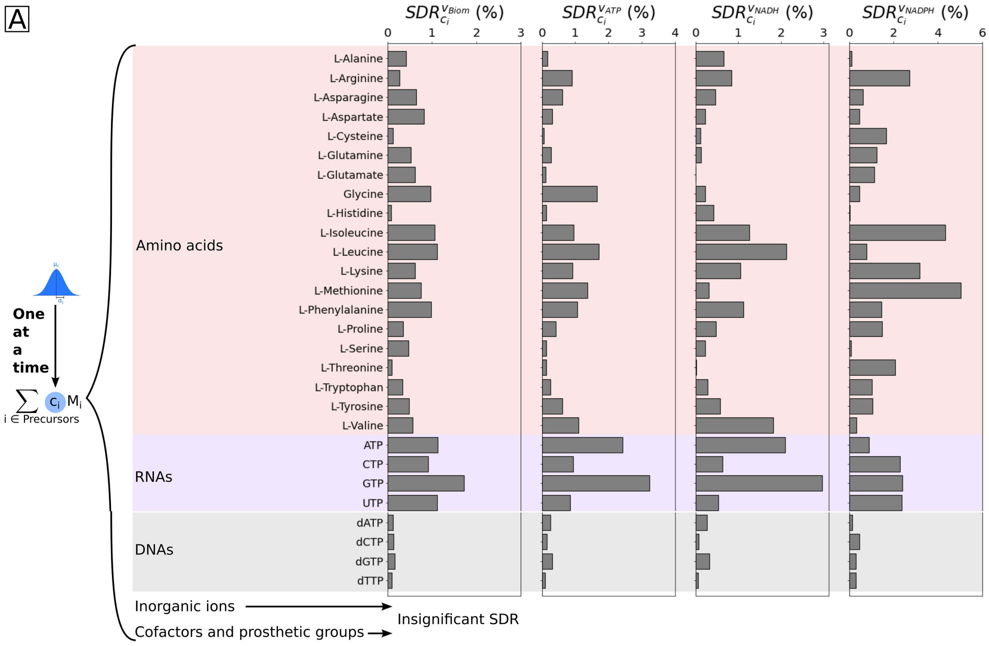


Figure 4.4: Results from Maranas et al. study on the impact of uncertainty injection into specific biomass consituents. It focused on the Standard Deviation Ratio between the injected uncertainty and the production of biomass, ATP, NADH and NADPH. Figure modified from Ref. [13]

varying, would not have a drastic impact on the growth of biomass as a whole. The results from the analysis performed on the iML1515, provided by Maranas et al. [13] study, can be seen in Fig. 4.4. By comparing the results from each model we can see that the eciML1515 generally displays the same overall patterns for growth rate and ATP production as the iML1515. However the iML1515 displays a consistently higher SDR for each individual biomass coefficient compared to the iML1515. This is consistent with expectations for the same reasons talked about in chapter (x). The eciML1515 has less degrees of freedoms owing to its additional set of enzyme restrictions, meaning it is less robust then the iML1515 and so is expected to display a higher SDR compared to the iML1515. However of note is also that the results for NADH- and NADPH production for the eciML1515 model are almost the same. The specific values of each one can be found in the Github repository. This is not shared between the results for iML1515, which have significant differences between NADH- and NADPH production. The reason why the eciML1515 has such similar results is unclear and still not fully understood. It seems like an unlikely result owing to the fact that NADPH is produced from NADH. These results seems to indicate that almost the entirety of NADH is converted into NADPH in the system, which seems like an unlikely result. It is possible that this is caused by the earlier mentioned pathways being disabled and causing a huge shift in the metabolic phenotype. This could have led to the pathways consuming NADH, for other purposes then to convert it to NADPH, being unavailable. Another possibility is that this is due to the source code to perform this step was not available. As such, it had to be implemented based on the information available in Maranas et al. study [13]. It is possible that this implementation was not done correctly and therefore the NADH, ATP and NADPH are displaying the wrong values. In order to evaluate this step could be re-performed after the eciML1515 has been retuned with the pathways disabled and with the code to perform this step re-evaluated. However, this is beyond the scope of this project.

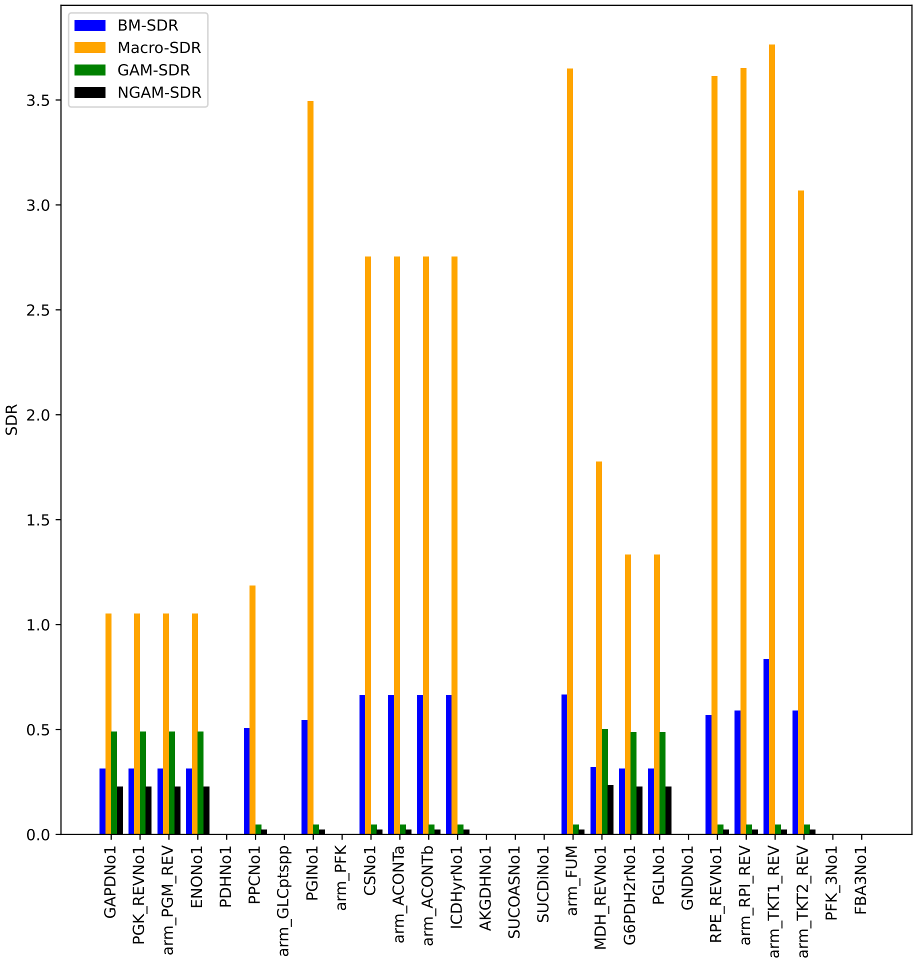


Figure 4.5: Propagation of uncertainty, focused on the citric acid cycle and glycolysis, from the parametric uncertainty injection into all biomass coefficents, macro-molecular fractions, GAM and NGAM. From this we can see that the macro results displayed an ampliyfing effect, while the other parameters displayed a dampening effect on the specific reactions, although the BM-SDR displayed a relatively minor dampening effect.

4.1.4 Propagation of uncertainty on specific fluxes in the glycolysis and citric acid cycle pathways from different parameters

After the parametric uncertainty on individual biomass coefficients was finished the project moved on to studying how the parametric uncertainty impacted the predicted internal fluxes of the metabolic network of eciML1515. The citric acid cycle and glycolysis was selected for closer examination as the study by Maranas et al. selected these pathways, giving a reference for comparison for the eciML1515. Additionally these pathways are part of the high-flux backbone and studies have shown that *E. coli* responds to perturbations in the growth conditions by reorganizing the pathways the rates of the fluxes that are part of this backbone [57]. By studying these pathways, the results should give insight into how the model responds to biomass composition perturbations. The SDR values of the reactions belonging to these pathways were thus extracted from the results. A graphical representation of the SDR of these reactions for the different parameters for uncertainty injection can be seen in Fig. 4.5.

From these results we can see that the macro uncertainty displays a massive amplifying effect on the uncertainty propagation throughout the predicted metabolic network for glycolysis and citric acid cycle. This reflects the earlier result where the uncertainty on the biomass production was amplified for the Macro-parameter for injection. It is expected that the Macro parameter displays an overall higher SDR compared to the all biomass coefficients parameter, owing to the fact that adding uncertainty into the macro molecular fractions of the biomass composition will naturally prioritize the elements of the biomass that are the most abundant. However, the size of the difference reflected in these results is unexpected. If we compare to the results from the iML1515, from Maranas et al. study, we can see that the results from the iML1515 shows Macro results for each reaction that never go above twice the results of the all biomass coefficients result. This is in stark comparison to the results from the eciML1515 that display Macro results up to five times higher compared to the biomass coefficients results. In addition the iML1515 model also displayed high results for the GAM parameter that are not shared for the eciML1515 model. These inconsistencies are possibly a result of the scripts for the Macro, GAM and NGAM parameters. While the study performed by Maranas et al. provided scripts for the injection of uncertainty for all the parameters, they only provided scripts for the calculating the sum of fluxes and SDR for the ”all biomass coefficients” parameter. As such the scripts for the other parameters had to be modified from the provided scripts. It is possible that the scrips were not modified correctly and does not calculate the correct values for the other parameters. Based on this, it was decided that only the results from the ”all biomass constituents” parameter was evaluated and these results can be seen in Fig. 4.6.

These results show that the uncertainty on the predicted internal metabolic fluxes from uncertainty injection of all biomass coefficients is different from the prediction of biomass production. While the biomass production displayed an SDR of about 0.43 from injection of uncertainty into all biomass coefficients, a number of the internal reactions display a higher SDR. While the reactions still display a dampening effect on propagation of uncertainty, it is lessened to the point that uncertainty of biomass composition can lead to different results on the internal flux distribution. This was expected as the robustness of the prediction of biomass growth is not reflected in the internal metabolic network of the model. The result is also shared for the iML1515 that displayed a similarly higher SDR for the internal reactions compared to the biomass production, as can be seen in the study by Maranas et al. [13].

Of note is that multiple reactions in the pathways displayed no results. This was unexpected as these reactions are connected and part of the same pathways, and therefore should display similar patterns of uncertainty propagation. In addition an SDR of 0 is a highly unlikely result for any active pathway in the metabolic network, as any active pathway will display some variance from pFBA. Therefore it is likely that these empty values point to some mistake in either the implementation of the scripts for calculating SDR or how the model interacts with the scripts.

In spite of these inconsistencies, it seems likely that the additional constraints from the enzyme constrained model had an impact on the propagation of uncertainty, however, a closer look at the model and the scripts should be performed in order to clear up the inconsistencies from the analysis.

These results indicate that the biomass compositions of the models are important for the metabolic flux predictions for the eciML1515, especially when studying the internal metabolic fluxes. This reinforces the results from the Ref. [13] analysis that showed a similar result for the internal metabolic flux predictions on the iML1515. Based on the result that perturbations in the biomass composition has an impact on the metabolic flux predictions of a model, espe-

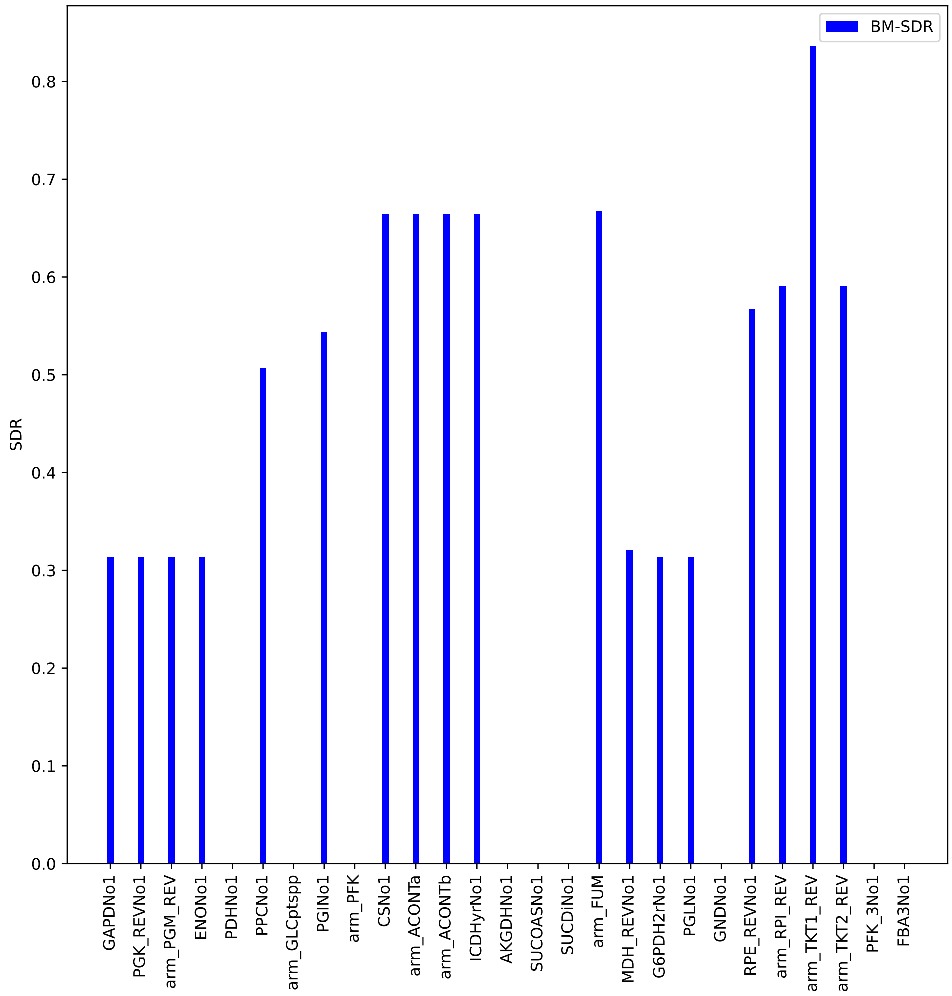


Figure 4.6: Propagation of uncertainty, focused on the citric acid cycle and glycolysis, from the parametric uncertainty injection into all biomass coefficents, represented by the BM-SDR. From this we can see that the macro results displayed an ampliyfing effect, while the other parameters displayed a dampening effect on the specific reactions, although the BM-SDR displayed a relatively minor dampening effect.

cially when additional constraints are implemented, one can intuit that the implementation of a BOF that more accurately represents the biomass composition under different environments and conditions could increase the predicted flux accuracy of the model. In order to analyse this further the next part of this project implemented growth rate-specific BOFs, in order to have a more accurate representation of the biomass composition, for an *E. coli* model and compared the predictions from these BOF variants with experimental data.

4.2 Quantifying the error and difference in flux predictions for growth rate-specific BOF variants of the iML1515 model compared to the standard variant

Owing to the impact on the predicted growth rate following the disabling of the reverse beta oxidase and reverse pyruvate synthase, shown in 4.1.1, it was decided that the eciML1515 would not be the model of choice for studying growth rate-specific BOFs. This is due to the difficulty it would have to reach the experimental values necessary for quantifying the error in flux predictions. Instead the standard iML1515 was used as a base mode for implementing growth rate-specific BOFs.

4.2.1 Implementing growth rate specific biomass objective functions based on experimental data

Eight different variations of the iML1515 model with different growth rate-specific biomass objective functions were created. The specific growth rates were based on experimental data provided by Gerosa et al. [15]. The models can be found in the Github repository under models. Each of the growth rate-specific BOF models were also bundled with a standard iML1515 model to act as a reference.

After the implementation of the growth rate-specific biomass objective functions, the bounds of the exchange reactions and growth rate were set according to the experimental data. However, this turned out to produce infeasible solutions for FBA on most of the models. This is likely due to the specific values from the experimental data not being in the available solution space for the iML1515 model. The iML1515 is a model of the *E. coli* strain K-12 MG1655 [58], while the experimental data is based on the *E. coli* BW25113 [15]. The difference in genomic data between these two can lead to the experimental data for one not being a feasible solution for the other.

As such it was decided that MOMA would be utilised in order to find the closest possible feasible solution to the specific growth rate and exchange reaction fluxes from the experimental data. However, COBRAPys MOMA function was found to be lacking in this regard as it uses the flux solution of a model as its reference values. Since, for this project, the reference values were the exchange fluxes and the growth rate the MOMA function provided by COBRAPy would not be applicable. The models were therefore ported over to ReFramed using its provided porting function and MOMA was utilised on the models as per 3.3.2.

After the MOMA was finished, the growth rate, secretion and uptake values were extracted from the results and set as new bounds for the models in COBRAPy. This allowed them to represent the closest feasible solution to the experimental data. However, if the values were set as bounds directly the models would still produce infeasible solutions. This is due to how

Python handles floats and rounding, as such the bounds had to be given a tiny amount of slack to account for Pythons rounding of floats.

After the bounds were properly implemented into each respective model, sampling of the models was performed as outlined in 3.4. The sampling produced 10 000 solutions with variations for each model. In order to evaluate the results, the absolute RMSE score for each solution was calculated using equation 3.20. This gives a comparison between the sampled fluxes and the experimental data. The flux for the reaction SUCOAS displayed a huge variance and would inflate the RMSE score for each sample massively. This is potentially a result of there being multiple reactions that form and consume a specific metabolite in equal measures, one of which being the SUCOAS reaction. This could create a metabolic loop of production and consumption of said metabolite, where these reactions can vary all the way up to the upper bound and still produce feasible metabolic phenotypes. Therefore, when sampling is performed, and the flux of SUCOAS is considered, the RMSE score appears largely inflated. To circumvent this inflation, and keep the RMSE scores sensible, a filter was added to remove the SUCOAS flux from consideration when calculating the RMSE scores.

This produced 10 000 RMSE scores from each sample of each model. In order to best evaluate these scores it was decided that a box plot representation would be made and a t-test would be performed for the reference models and the growth rate-specific BOF models. The box plot representation can be seen in Fig. 4.7 and the t-test results can be seen in Table 4.2.

From these results we can see that, by implementing growth rate-specific BOFs, the models show some improvements for galactose, fructose, pyruvate and glycerol. The standard iML1515 model was expected to display a better result then the reconstructed biomass objective function one for glucose as a carbon source, as the iML1515 was built based on data from a glucose limited *E.coli* cell [58].

After some discussion around these results it was decided that another statistical analysis for the results would be performed with relative error evaluation and a cutoff point instead. This was because the absolute RMSE score would overvalue fluxes with a large base value, causing an inflation of the RMSE score. On the other hand, a relative error with no cutoff would massively overvalue the tiny fluxes in the model, causing another inflation of RMSE score. By introducing a cutoff point, where only fluxes above a certain size are considered, this problem can be avoided and the relative error can still be evaluated. It is also worth noting that the absolute RMSE score ignores the standard deviation of the experimental flux data. However, from the data, provided in the supplementary data, it can be seen that some of these values have large standard deviation, meaning that the experimental value provided for the flux is unlikely to be an accurate representation. To account for this another filter was added where, if a reaction in the experimental data displayed an SD that was half of the value or above, a relative standard deviation of 50%, it would not be considered for the error evaluation.

For these reasons the relative error of each sample value was considered against the experimental data if the value was above 0.1 and the SD was not higher then half the value. A box plot comparing this relative error can be seen in Fig. 4.8 and a t-test comparison between the population of the growth rate-specific models and the standard model can be seen in Table 4.3.

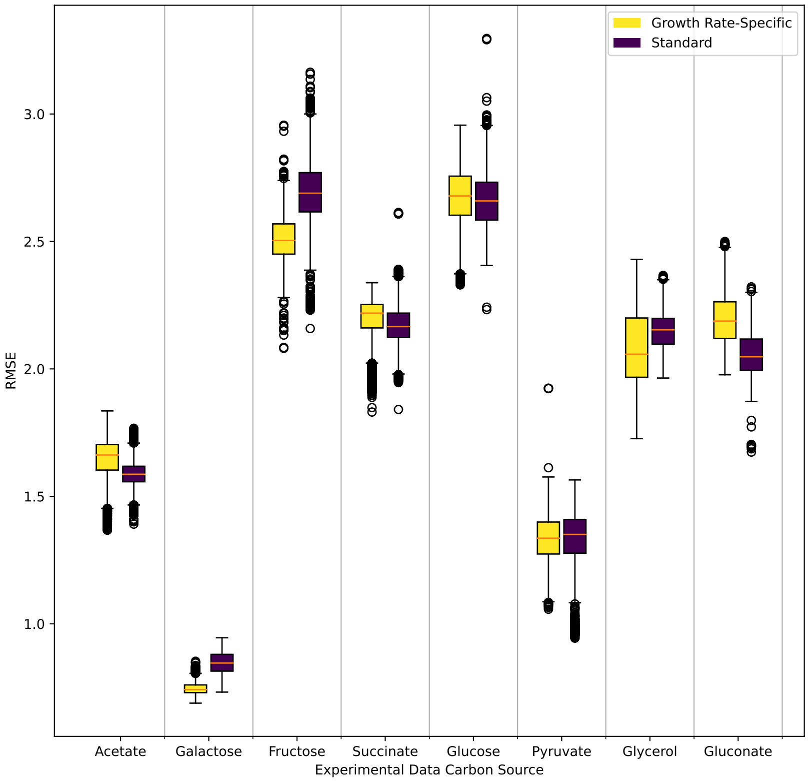


Figure 4.7: Box plot of the RMSE score between the growth rate-specific BOF variants and the reference BOF of the iML1515 model. The growth rate-specific BOF variants were implemented based on experimental data from different carbon sources, producing eight different variants. The carbon source for each is displayed on the x-axis. The bounds for each model were taken from a MOMA solution in order to find the closest bounds to the experimental data that still produced a feasible model. Each model was sampled, along with a reference model, to produce 10 000 different solution and each solution was compared with the experimental data for the internal fluxes in order to produce the RMSE score. The RMSE score is represented in the y-axis.

Table 4.2: Results from the t-test performed between the RMSE score for the growth rate-specific BOF variants and the reference iML1515 models. The growth rate-specific BOF variants were implemented based on experimental data from an *E. coli* strain growing on mediums with different carbon sources. The RMSE score was calculated based on sampling results, where an approximation of the experimental values for the exchange rates and growth rates were used as bounds for the models. We can see that all the models displayed a low p-value meaning that the difference in the populations are all statistically significant. In addition, the galactose, fructose, pyruvate and glycerol show a negative T-value confirming the results from Fig. 4.7 that the growth rate-specific BOF variant is more accurate to the experimental data compared to the standard iML1515 model for these carbon sources.

Carbon Source	T-value	p
Acetate	69.14	0.000
Galactose	-194.80	0.000
Fructose	-134.30	0.000
Succinate	18.90	0.000
Glucose	8.00	0.000
Pyruvate	-2.97	0.003
Glycerol	-46.28	0.000
Gluconate	102.20	0.000

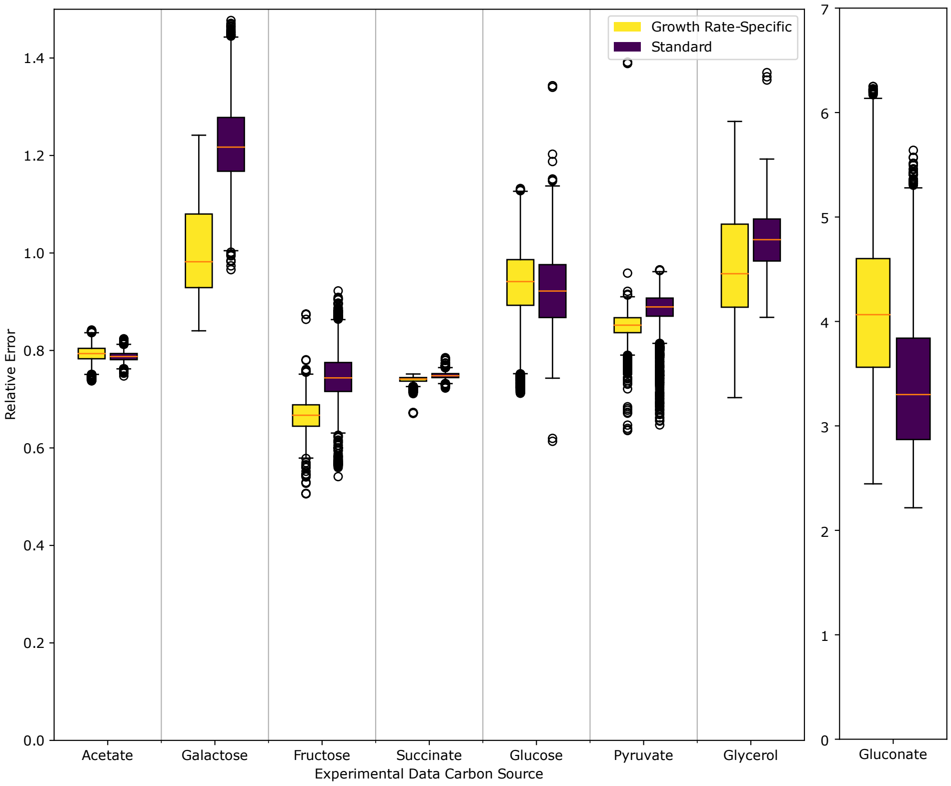


Figure 4.8: Box plot of the relative error of the growth rate-specific BOF variants and the reference BOF of the iML1515 model. The growth rate-specific BOF variants were implemented based on experimental data from different carbon sources, producing eight different variants. The carbon source for each is displayed on the x-axis. The bounds for each model were taken from a MOMA solution in order to find the closest bounds to the experimental data that still produced a feasible model. Each model was sampled, along with a reference model, to produce 10 000 different solution and each solution was compared with the experimental data for the internal fluxes in order to produce the relative error. In addition a filter was implemented that removed values that were either ≤ 0.1 in flux value or had an SD that was $\geq 50\%$ of the flux value. The relative error is represented in the y-axis.

Table 4.3: Results from the t-test performed between the relative error for the growth rate-specific BOF variants and the reference iML1515 models. The growth rate-specific BOFs were implemented based on experimental data from *E. coli* strain growing on media with different carbon sources. The relative error were calculated based on sampling results, where an approximation of the experimental values for the exchange rates and growth rates were used as bounds for the models. In addition a filter was implemented that removed values that were either ≤ 0.1 in flux value or had an SD that was $\geq 50\%$ of the flux value. As can be seen all the models displayed a p-value of 0.000, meaning that the populations are statistically significantly different from each other. In addition the data based on galactose, fructose, succinate, pyruvate and glycerol as carbon sources all displayed negative T-values. This means that the values for the growth rate-specific BOFs variants displayed a lower relative error then the values from the standard reference iML1515 models.

Carbon Source	T-value	p
Acetate	31.43	0.000
Galactose	-174.77	0.000
Fructose	-147.63	0.000
Succinate	-94.36	0.000
Glucose	12.42	0.000
Pyruvate	-87.48	0.000
Glycerol	-47.06	0.000
Gluconate	75.68	0.000

The results from the relative error comparison reinforces the results from the absolute comparison. Showing improvements in the predicted flux accuracy for galactose, fructose, pyruvate and glycerol. In addition, the relative error showed improvements for the experimental data based on the succinate carbon source. This is likely a result of the growth rate-specific BOF variants having more accurate results, for the experimental data, for the smaller fluxes. This would mean that the relative values display results closer to 0, which further means the results are closer to the experimental data, for the growth rate-specific BOF variants compared to the standard iML1515. This is reflected in the results and points to the attempt at implement a growth rate-specific BOF was successful.

This could mean that the implementation of more condition specific BOFs has potential for increasing the predicted flux accuracy of metabolic models. The results seems to indicate that the growth rate-specific BOF achieved a more accurate prediction of metabolic fluxes compared to the standard BOF. The growth rate-specific BOF displayed values that were closer to 0, when looking at the relative error, for the galactose, fructose, succinate, pyruvate and glycerol data sets. As mentioned earlier the standard BOF for the iML1515 is based on data from *E. coli* strains mainly from glucose as a carbond source, as such it was unlikely that the growth rate-specific BOF would achieve a more accurate prediction then the standard one. This is also reflected in the relative error results, where the standard BOF displayed a relative error that was closer to 0. Nevertheless, by tailoring the biomass composition to be more specific to certain conditions and strains, it seems possible that more accurate predictions of metabolic phenotypes can be achieved.

This data also indicates that the accuracy of the GEM predictions for the internal metabolic fluxes are poor. The relative error of each carbon source was grouped between 0.8 and 1.2 for each carbon source, with the exception of gluconate which was at around 3.5 relative error. A relative error of 1 means that the average samples average predicted flux was a full value off the experimental data. This indicates that the GEMs, in their current state, perform poorly for predictions of the internal metabolic fluxes and studies should be performed in order to find ways to improve the GEMs for these aspects.

It should be noted that this data is based on experimental values that displayed high values of standard deviation. As mentioned earlier steps were taken to filter the values that displayed too much variation, or that were too low to be considered significant. However, the limit for when these values were filtered away and not compared to the flux predictions were arbitrarily chosen. It is possible that more accurate limits for filtration could improve the data and the results, but this would require an in depth consideration of each reaction and would be above the time budget available for this project.

Another way to avoid this issue could be to gather more specific data, which does not display as large deviation. However, this is also beyond the scope of this project. Additionally, this study only looked at the fluxes tied to pathways in the central carbon metabolism, which should be relatively robust. Some other specific pathways in the metabolism, that are more directly tied to the production of biomass components, are likely to display a larger difference between the growth rate-specific BOF and the standard BOF. As an example, the synthesisation of nucleotides is directly tied to the production of RNA in a cell, and therefore should be dependent on their fraction in the biomass composition. However, the data found on these pathways was limited and not focused on the growth rate and therefore would not be a good fit for this study.

Also of note is that the experimental data, the iML1515 model and the growth rate-specific biomass data are all based on different strains of *E. coli*. It is possible that more specific data focusing on one strain could further improve the accuracy. Although this would involve gathering specific data for the strain and is an undertaking beyond the scope of this project.

Chapter 5

Conclusion and Outlook

This thesis was an attempt to further the understanding of GEM models internal flux distribution and how additional constraints and variables impact this accuracy. The first part of the study focuses on the propagation of uncertainty in an enzyme constrained model. Specifically the *E. coli* GEM eciML1515. The theoretical foundation was that the additional constraints added with implementation of enzymatic data would give a more accurate view of the impact of uncertainty on flux predictions from perturbations in biomass compositions. This is due to how the eciML1515 is naturally less robust and likely a better representation of an *E. coli* cell in this regard compared to the iML1515.

To compare the propagation of uncertainty analysis results between the models, data from Maranas et al. study was collected while a similar analysis was attempted for the eciML1515. The eciML1515 displayed a significantly lowered level of robustness from this analysis compared to the standard GEM model of iML1515. This supports the theoretical foundation by illustrating that the additional constraints and lowered degrees of freedom could mean that uncertainty in the biomass composition has a larger impact on flux predictions. There were some inconsistencies with the other parameters, which could have impacted the results and the conclusions of these parameters. With more resources and time the eciML1515 could be retuned and improved, as well as the code for the perturbation analysis. However such endeavours are beyond the scope of this project.

The second part of this project aimed to study how the implementation of a growth rate-specific BOF would improve the predicted flux distribution for GEM models. The theoretical foundation for this part was since the biomass composition has an impact on metabolic flux predictions, as the results of the Ref. [13] analysis and the earlier work in this project illustrated, a more accurate representation of the composition would translate to a more accurate flux prediction. In order to study this, experimental data was used to create eight different growth rate-specific BOF variants. The experimental data was based on samples collected from *E. coli* cells collected from growth media with differing carbon sources. These variants were then compared to the standard iML1515. The growth rate-specific variants showed improvements in predictions for the majority of the carbon sources, with the exception of acetate, glucose and gluconate. The variants were expected to not show much improvement for glucose, as the iML1515 is based on data collected from an *E.coli* growing with a glucose carbon source.

This shows that the implementation of condition-dependent biomass based on growth rate can improve predicted flux distributions for models and is an avenue of models that should be further explored. In particular the impact on specific pathways in the cell should be studied closer, to better understand how growth rate effects the metabolic flux network. Additionally, studies on other conditions that could be implemented in order to improve the accuracy of the current GEM models could be studied closer.

Bibliography

[1] Sabina Leonelli. *Data-Centric Biology*. University of Chicago Press, Chicago, 2016. ISBN 9780226416502. doi: doi:10.7208/9780226416502. URL `https://doi.org/10.7208/9780226416502`.

[2] Francis S Collins and Leslie Fink. The human genome project. *Alcohol Health Res. World*, 19(3):190–195, 1995.

[3] Mcdonnell genome institute (mgi). URL `https://genome.wustl.edu/`. Accessed: 2023-04-25.

[4] Alon Bartal and Kathleen M. Jagodnik. Progress in and opportunities for applying information theory to computational biology and bioinformatics. *Entropy*, 24(7), 2022. ISSN 1099-4300. doi: 10.3390/e24070925. URL `https://www.mdpi.com/1099-4300/24/7/925`.

[5] Eman Karam ElSayed, Iman Ahmed ElSayed, Kamal ElDahshan, and Hesham Hefny. Big data and its future in computational biology: A literature review. *Journal of Computer Science*, 17:1222–1228, 2021.

[6] S. ć. Systems biology, emergence and antireductionism. *Saudi J Biol Sci*, 23(5):584–591, Sep 2016.

[7] Tong Ihn Lee, Nicola J. Rinaldi, François Robert, Duncan T. Odom, Ziv Bar-Joseph, Georg K. Gerber, Nancy M. Hannett, Christopher T. Harbison, Craig M. Thompson, Itamar Simon, Julia Zeitlinger, Ezra G. Jennings, Heather L. Murray, D. Benjamin Gordon, Bing Ren, John J. Wyrick, Jean-Bosco Tagne, Thomas L. Volkert, Ernest Fraenkel, David K. Gifford, and Richard A. Young. Transcriptional regulatory networks in ¡i¿saccharomyces cerevisiae¡/i¿. *Science*, 298(5594):799–804, 2002. doi: 10.1126/science.1075090. URL `https://www.science.org/doi/abs/10.1126/science.1075090`.

[8] Avi Ma'ayan. Introduction to network analysis in systems biology. *Sci. Signal.*, 4(190): tr5, September 2011.

[9] Anurag Passi, Juan D Tibocha-Bonilla, Manish Kumar, Diego Tec-Campos, Karsten Zengler, and Cristal Zuniga. Genome-scale metabolic modeling enables in-depth understanding of big data. *Metabolites*, 12(1):14, December 2021.

[10] Won Jun Kim, Hyun Uk Kim, and Sang Yup Lee. Current state and applications of microbial genome-scale metabolic models. *Current Opinion in Systems Biology*, 2:10–18, 2017. ISSN 2452-3100. doi: https://doi.org/10.1016/j.coisb.2017.03.001. URL https://www.sciencedirect.com/science/article/pii/S2452310017300483. Regulatory and metabolic networks • Cancer and systemic diseases.

[11] Adam M Feist and Bernhard O Palsson. The biomass objective function. *Curr. Opin. Microbiol.*, 13(3):344–349, June 2010.

[12] J. Pramanik and J. D. Keasling. Stoichiometric model of escherichia coli metabolism: Incorporation of growth-rate dependent biomass composition and mechanistic energy requirements. *Biotechnology and Bioengineering*, 56(4):398–421, 1997. doi: https://doi.org/10.1002/(SICI)1097-0290(19971120)56:4⟨398::AID-BIT6⟩3.0.CO;2-J.

[13] Costas D. Maranas, Hoang V. Dinh, and Debolina Sarkar. Quantifying the propagation of parametric uncertainty on flux balance analysis. *Metabolic Engineering*, 69:26–39, 2022. ISSN 1096-7176. doi: https://doi.org/10.1016/j.ymben.2021.10.012. URL https://www.sciencedirect.com/science/article/pii/S1096717621001634.

[14] Philip L. Bonner Trevor Palmer. *Enzymes: Biochemistry, Biotechnology and Clinical Chemistry, 2nd edition*, chapter 6 An Introduction to Bioenergetics, Catalysis and Kinetics. Woodhead Publishing Limited, 2007.

[15] Luca Gerosa, Bart R.B. Haverkorn van Rijsewijk, Dimitris Christodoulou, Karl Kochanowski, Thomas S.B. Schmidt, Elad Noor, and Uwe Sauer. Pseudo-transition analysis identifies the key regulators of dynamic metabolic adaptations from steady-state data. *Cell Systems*, 1(4):270–282, Oct 2015. ISSN 2405-4712. doi: 10.1016/j.cels.2015.09.008. URL https://doi.org/10.1016/j.cels.2015.09.008.

[16] Philip L. Bonner Trevor Palmer. *Enzymes: Biochemistry, Biotechnology and Clinical Chemistry, 2nd edition*, chapter 7 Kinetics of Single-Substrate EnzymeCatalysed Reactions. Woodhead Publishing Limited, 2007.

[17] C. et al. Rye. 6.1 energy and metabolism. https://openstax.org/books/biology/pages/6-1-energy-and-metabolism. Accessed: 2023-4-20.

[18] Joanne M. Savinell and Bernhard O. Palsson. Network analysis of intermediary metabolism using linear optimization. i. development of mathematical formalism. *Journal of Theoretical Biology*, 154(4):421–454, 1992. ISSN 0022-5193. doi: https://doi.org/10.1016/S0022-5193(05)80161-4. URL https://www.sciencedirect.com/science/article/pii/S0022519305801614.

[19] Dongsoo Yang, Seon Young Park, Yae Seul Park, Hyunmin Eun, and Sang Yup Lee. Metabolic engineering of escherichia coli for natural product biosynthesis. *Trends in Biotechnology*, 38(7):745–765, July 2020. doi: 10.1016/j.tibtech.2019.11.007. URL https://doi.org/10.1016/j.tibtech.2019.11.007.

[20] Adam M Feist, Markus J Herrgård, Ines Thiele, Jennie L Reed, and Bernhard Ø Palsson. Reconstruction of biochemical networks in microorganisms. *Nat. Rev. Microbiol.*, 7(2): 129–143, February 2009.

[21] Bernhard Palsson. Metabolic systems biology. *FEBS Letters*, 583(24): 3900–3904, 2009. doi: https://doi.org/10.1016/j.febslet.2009.09.031. URL https://febs.onlinelibrary.wiley.com/doi/abs/10.1016/j.febslet.2009.09.031.

[22] Bernhard O. Palsson. *Systems Biology: Constraint Based Reconstruction and Analysis*, chapter 17 Constraints. Cambride University Press, 2015.

[23] Bernhard O. Palsson. *Systems Biology: Constraint Based Reconstruction and Analysis*, chapter 9 The Stoichiometric Matrix. Cambride University Press, 2015.

[24] Hendrik P.J. Bonarius, Georg Schmid, and Johannes Tramper. Flux analysis of underdetermined metabolic networks: the quest for the missing constraints. *Trends in Biotechnology*, 15(8):308–314, August 1997. doi: 10.1016/s0167-7799(97)01067-6. URL https://doi.org/10.1016/s0167-7799(97)01067-6.

[25] Amit Varma and Bernhard O. Palsson. Metabolic flux balancing: Basic concepts, scientific and practical use. *Bio/Technology*, 12(10):994–998, October 1994. doi: 10.1038/nbt1094-994. URL https://doi.org/10.1038/nbt1094-994.

[26] Mark A Schulze. Linear programming for optimization. *Perspective Scientific Instruments Inc., USA*, 1998.

[27] Mark Schulze. Linear programming for optimization. 09 2000.

[28] Hennie de Harder. A beginner's guide to linear programming and the simplex algorithm, Jan 2023. URL https://towardsdatascience.com/a-beginners-guide-to-linear-programming-and-the-simplex-algorithm-87d

[29] Steffen Klamt and Axel von Kamp. Analyzing and resolving infeasibility in flux balance analysis of metabolic networks. *Metabolites*, 12(7):585, June 2022.

[30] Jeffrey D. Orth, Ines Thiele, and Bernhard Ø Palsson. What is flux balance analysis? *Nature Biotechnology*, 28(3):245–248, Mar 2010. ISSN 1546-1696. doi: 10.1038/nbt.1614. URL https://doi.org/10.1038/nbt.1614.

[31] Bernhard O. Palsson. *Systems Biology: Constraint Based Reconstruction and Analysis*, chapter 18 Optimization. Cambride University Press, 2015.

[32] Rafael U. Ibarra, Jeremy S. Edwards, and Bernhard O. Palsson. Escherichia coli k-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature*, 420 (6912):186–189, November 2002. doi: 10.1038/nature01149. URL https://doi.org/10.1038/nature01149.

[33] Jeffrey D. Orth, Ines Thiele, and Bernhard Ø Palsson. What is flux balance analysis? *Nature Biotechnology*, 28(3):245–248, Mar 2010. ISSN 1546-1696. doi: 10.1038/nbt.1614. URL https://doi.org/10.1038/nbt.1614.

[34] Carlos Eduardo García Sánchez and Rodrigo Gonzalo Torres Sáez. Comparison and analysis of objective functions in flux balance analysis. *Biotechnology Progress*, 30 (5):985–991, 2014. doi: https://doi.org/10.1002/btpr.1949. URL https://aiche.onlinelibrary.wiley.com/doi/abs/10.1002/btpr.1949.

[35] Kenneth J Kauffman, Purusharth Prakash, and Jeremy S Edwards. Advances in flux balance analysis. *Current Opinion in Biotechnology*, 14(5):491–496, October 2003. doi: 10.1016/j.copbio.2003.08.001. URL `https://doi.org/10.1016/j.copbio.2003.08.001`.

[36] Maximilian Lularevic, Andrew J. Racher, Colin Jaques, and Alexandros Kiparissides. Improving the accuracy of flux balance analysis through the implementation of carbon availability constraints for intracellular reactions. *Biotechnology and Bioengineering*, 116(9): 2339–2352, June 2019. doi: 10.1002/bit.27025. URL `https://doi.org/10.1002/bit.27025`.

[37] Nathan E Lewis, Kim K Hixson, Tom M Conrad, Joshua A Lerman, Pep Charusanti, Ashoka D Polpitiya, Joshua N Adkins, Gunnar Schramm, Samuel O Purvine, Daniel Lopez-Ferrer, Karl K Weitz, Roland Eils, Rainer König, Richard D Smith, and Bernhard Ø Palsson. Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. *Molecular Systems Biology*, 6(1):390, January 2010. doi: 10.1038/msb.2010.47. URL `https://doi.org/10.1038/msb.2010.47`.

[38] Daniel Machado and Markus Herrgård. Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. *PLoS Computational Biology*, 10(4):e1003580, April 2014. doi: 10.1371/journal.pcbi.1003580. URL `https://doi.org/10.1371/journal.pcbi.1003580`.

[39] A Varma, B W Boesch, and B O Palsson. Stoichiometric interpretation of escherichia coli glucose catabolism under various oxygenation rates. *Applied and Environmental Microbiology*, 59(8):2465–2473, August 1993. doi: 10.1128/aem.59.8.2465-2473.1993. URL `https://doi.org/10.1128/aem.59.8.2465-2473.1993`.

[40] Daniel Segrè, Dennis Vitkup, and George M. Church. Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences*, 99(23):15112–15117, November 2002. doi: 10.1073/pnas.232349399. URL `https://doi.org/10.1073/pnas.232349399`.

[41] Adam M Feist and Bernhard O Palsson. The biomass objective function. *Current Opinion in Microbiology*, 13(3):344–349, June 2010. doi: 10.1016/j.mib.2010.03.003. URL `https://doi.org/10.1016/j.mib.2010.03.003`.

[42] Vetle Simensen, Christian Schulz, Emil Karlsen, Signe Bråtelund, Idun Burgos, Lilja Brekke Thorfinnsdottir, Laura García-Calvo, Per Bruheim, and Eivind Almaas. Experimental determination of escherichia coli biomass composition for constraint-based metabolic modeling. *PLOS ONE*, 17(1):e0262450, January 2022. doi: 10.1371/journal.pone.0262450. URL `https://doi.org/10.1371/journal.pone.0262450`.

[43] Jonathan M Monk, Colton J Lloyd, Elizabeth Brunk, Nathan Mih, Anand Sastry, Zachary King, Rikiya Takeuchi, Wataru Nomura, Zhen Zhang, Hirotada Mori, Adam M Feist, and Bernhard O Palsson. iML1515, a knowledgebase that computes escherichia coli traits. *Nature Biotechnology*, 35(10):904–908, October 2017. doi: 10.1038/nbt.3956. URL `https://doi.org/10.1038/nbt.3956`.

[44] Frederick C Neidhardt and etc., editors. *Escherichia coli and salmonella typhimurium: Vols 1-2*. American Society for Microbiology, Washington, D.C., DC, January 1987.

[45] Benjamín J Sánchez, Cheng Zhang, Avlant Nilsson, Petri-Jaan Lahtvee, Eduard J Kerkhoven, and Jens Nielsen. Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Molecular Systems Biology*, 13(8):935, August 2017. doi: 10.15252/msb.20167411. URL `https://doi.org/10.15252/msb.20167411`.

[46] KR Srinath. Python–the fastest growing programming language. *International Research Journal of Engineering and Technology*, 4(12):354–357, 2017.

[47] Anaconda software distribution, 2020. URL `https://docs.anaconda.com/`.

[48] Ali Ebrahim, Joshua A Lerman, Bernhard O Palsson, and Daniel R Hyduke. COBRApy: COnstraints-based reconstruction and analysis for python. *BMC Systems Biology*, 7(1), August 2013. doi: 10.1186/1752-0509-7-74. URL `https://doi.org/10.1186/1752-0509-7-74`.

[49] Daniel Machado. cdanielmachado/reframed: 1.3.0, March 2023. URL `https://doi.org/10.5281/zenodo.7701852`.

[50] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2023. URL `https://www.gurobi.com`.

[51] SysBioChalmer. Sysbiochalmer enzyme constrained models, 2020. URL `https://github.com/SysBioChalmers/ecModels`.

[52] Katia Tarasava, Seung Hwan Lee, Jing Chen, Michael Köpke, Michael C Jewett, and Ramon Gonzalez. Reverse β-oxidation pathways for efficient chemical production. *J. Ind. Microbiol. Biotechnol.*, 49(2), April 2022.

[53] Takayuki Nakayama, Shin-Ichiro Yonekura, Shuji Yonei, and Qiu-Mei Zhang-Akiyama. ¡i¿escherichia coli¡/i¿ pyruvate:flavodoxin oxidoreductase, ydbk - regulation of expression and biological roles in protection against oxidative stress. *Genes Genetic Systems*, 88(3): 175–188, 2013. doi: 10.1266/ggs.88.175.

[54] J. Pramanik and J. D. Keasling. Stoichiometric model ofEscherichia coli metabolism: Incorporation of growth-rate dependent biomass composition and mechanistic energy requirements. *Biotechnology and Bioengineering*, 56(4):398–421, November 1997. doi: 10.1002/(sici)1097-0290(19971120)56:4⟨398::aid-bit6⟩3.0.co;2-j. URL `https://doi.org/10.1002/(sici)1097-0290(19971120)56:4<398::aid-bit6>3.0.co;2-j`.

[55] Veronica W. Rowlett, Venkata K. P. S. Mallampalli, Anja Karlstaedt, William Dowhan, Heinrich Taegtmeyer, William Margolin, and Heidi Vitrac. Impact of membrane phospholipid alterations in escherichia coli on cellular function and bacterial stress adaptation. *Journal of Bacteriology*, 199(13), July 2017. doi: 10.1128/jb.00849-16. URL `https://doi.org/10.1128/jb.00849-16`.

[56] Katia Tarasava, Seung Hwan Lee, Jing Chen, Michael Köpke, Michael C Jewett, and Ramon Gonzalez. Reverse -oxidation pathways for efficient chemical production. *Journal of Industrial Microbiology and Biotechnology*, 49(2), 02 2022. ISSN 1367-5435. doi: 10.1093/jimb/kuac003. URL `https://doi.org/10.1093/jimb/kuac003`. kuac003.

[57] E. Almaas, B. Kovács, T. Vicsek, Z. N. Oltvai, and A.-L. Barabási. Global organization of metabolic fluxes in the bacterium escherichia coli. *Nature*, 427(6977):839–843, February 2004. doi: 10.1038/nature02289. URL `https://doi.org/10.1038/nature02289`.

[58] Jonathan M Monk, Colton J Lloyd, Elizabeth Brunk, Nathan Mih, Anand Sastry, Zachary King, Rikiya Takeuchi, Wataru Nomura, Zhen Zhang, Hirotada Mori, Adam M Feist, and Bernhard O Palsson. iML1515, a knowledgebase that computes escherichia coli traits. *Nat. Biotechnol.*, 35(10):904–908, October 2017.

Appendix A

Data and Source Code

The data for the results of, as well as the source code for the methods used, in this project can be found in the Github repository at the URL:
`https://github.com/HelgeMonsson/MasterThesis`

NTNU

Norwegian University of
Science and Technology