Towards the Determination of the Active Set of Elementary Flux Modes using Metabolic Flux Data

Koren Murphy

Doctor of Philosophy

Faculty of Science, Agriculture and Engineering
School of Engineering
Chemical Engineering

Submission: 09/2023

Abstract

Vaccine development at lab scale through to large scale production can take 10-15 years. With the outbreak of the SARS-CoV-2 disease, emphasis on fast vaccine production was emphasised. However, the cells that are grown to produce antigens have complex metabolic networks consisting of thousands of reactions, metabolites, and genes. There is little understanding of why a cell in the same environmental conditions may grow via one route over another. If this process was better analysed, process optimisation to increase biomass growth and reduce inhibiting metabolites could be performed. All routes that a cell can use during its life are collectively known as elementary flux modes. Genome networks are being constructed over time allowing for full reaction stoichiometry to be known. However, genome networks do not have all the elementary flux modes identified due to the combinatorial explosion that occurs when solving as there can be billions of possible routes.

In this thesis, mixed integer linear programming has been presented to enumerate elementary flux modes as a future proof method towards genome scale solving. It is compared to publicly available tools and mixed integer linear programming methods throughout literature. The benefits of this method in the future for finding elementary flux modes are also discussed. Compression techniques and code parallelisation are examined and reduced solve times presented. Alongside elementary flux mode enumeration this thesis also applies flux analysis techniques as a method of finding biologically relevant elementary flux modes. Disadvantages of these techniques are highlighted whilst presenting an integrated form of metabolic flux analysis to alleviate some of the issues. The technique presented is proven to be a viable method for enumerating elementary flux modes with the integration of fluxes.

E.coli can be modelled as the full genome network or a reduced set of reactions representing the key areas of the network; this is known as the core network. *E. coli* fermentation data from GlaxoSmithKline was provided for this work, allowing for analysis techniques identified and created in this work to be applied. However, this data was found to be underdetermined preventing aspects to flux analysis and elementary flux mode enumeration to be performed. This thesis discusses the process data and estimates specific growth and uptake rates for all

fermentations in batch and fed batch operations. This key data was missing and helps in better understanding the operations taking place in the fermenters. More importantly however areas where more data is required for flux analysis are presented along with the issues of data limitation on finding the elementary flux modes even for the core network. Underdetermined flux analysis allowed for estimations on the number of possible elementary flux modes in batch and fed batch operations, highlighting the reduction in feasible routes during fed batch due to the cell's phase.

Acknowledgements

Firstly, thank you to my supervisors Dr Mark Willis and Dr Chris O'Malley. Embarking on this PhD was tough but having great knowledge behind me was paramount to me getting to his point. Your patience and willingness to have a laugh over the last four years has been invaluable. Secondly, to Moritz, Gael, Anne, Rui, and Sergio, you helped me delve into an industry that I only knew by name. Your willingness to sit down and teach me basics, share ideas, and teach me about your careers has been instrumental.

To my mum, dad, and sisters you have all been a listening ear to my complaints over the last few years and have offered your support whenever I needed it. To all my friends you have given me the much need breaks from work whilst always cheering me on. Although he won't be able to read this, I would like to thank my dog Poirot for getting me away from the computer screen every day which gave me time to come up with new ideas for this work.

And finally, the biggest thank you goes to Josh. After starting our undergraduate degrees together to both now finishing 8 years on engaged it's been invaluable to have you by my side. No question was ever too stupid, and every bad day was always made better by you. Thank you.

Table of Contents

List of Figures

List of Tables

Nomenclature

$\boldsymbol{A_R}$	Reaction centric adjacency matrix
$\boldsymbol{A_M}$	Metabolite centric adjacency matrix
$\boldsymbol{A_{M,R}}$	Bipartite adjacency matric
a	Maximum rate achieved by the system
a_i	Maximum specific uptake/excretion rate
$\boldsymbol{B}^{(m)}$	Reversible reaction tableau section
b_{ij}	Element in reversible reaction tableau
$\boldsymbol{c}$	Concentration
c_i	Concentration of metabolite 'i'
$\boldsymbol{c_F}$	Concentration of metabolites in the feed
$\boldsymbol{E_i}$	Elementary flux mode
$\boldsymbol{e}$	Extreme vector/ray
e	Number of EFMs
$\boldsymbol{F}^{(m)}$	Irreversible reaction tableau section
F	Feed rate
f_{ij}	Element in irreversible reaction tableau
G	Gene concentration
$\boldsymbol{I}$	Identity matrix
J	Cost function
K	Michaelis constant
$\boldsymbol{K}$	Matrix of linearly independent column vectors to define dependencies in the columns of $\boldsymbol{S_i}$
$\boldsymbol{K_r}$	Kernel
M_i	Big M constant
$\boldsymbol{m}$	Mass
m_{ex}	Number of extracellular metabolites
N_m	Total number of metabolites
N_p	Total number of products
N_R	Number of reactions

N_{rev}	Number of reversible reactions
N_s	Total number of substrates
r	Reaction
R	Reaction after compression
$\boldsymbol{S}$	Total stoichiometry
$\boldsymbol{S_e}$	Stoichiometry containing extracellular metabolites only
$\boldsymbol{S_{EFM}}$	Stoichiometry containing EFM reactions only
$\boldsymbol{S_i}$	Stoichiometry containing intracellular metabolites only
$\boldsymbol{S_{irr}}$	Irreversible reaction stoichiometry
$\boldsymbol{S_K}$	Macroscopic stoichiometry
$\boldsymbol{S_m}$	Stoichiometry of measured variables
$\boldsymbol{S_N}$	Stoichiometry consisting of the reactions with unknown fluxes
$\boldsymbol{S_P}$	Product stoichiometry
$\boldsymbol{S_p}$	Product stoichiometric numbers
$\boldsymbol{S_r}$	Reactant stoichiometric numbers
$\boldsymbol{S_{rev}}$	Reversible reaction stoichiometry
$\boldsymbol{S_S}$	Substrate stoichiometry
$\boldsymbol{T}^{(m)}$	Double descriptive tableau
t	Time
u	Unknowns
V	Volume
$\boldsymbol{v}$	Flux vector
v	Specific uptake rate
v_i	Specific reaction flux
$\boldsymbol{v_{irr}}$	Irreversible reaction flux vector
$\boldsymbol{v_m}$	Measured flux vector
$v_{m,min}$	Minimum measured flux vector
$v_{m,max}$	Maximum measured flux vector
$\boldsymbol{v_{rev}}$	Reversible reaction flux vector
w	EFM weighting
X	Biomass concentration

Z	Objective function
$\boldsymbol{z}$	Artificial variables for error minimisation
δ_i	Binary variable
$\delta_{f,l}$	Binary variable of forward reaction
$\delta_{r,l}$	Binary variable of reverse reaction
$\delta_{s,m}$	Substrate binary variables
$\delta_{p,n}$	Product binary variables
$\delta_{c,1}$	Binary variable of consuming reaction
$\delta_{pr,1}$	Binary variable of producing reaction
μ	Specific growth rate
$\boldsymbol{\mu}$	Weightings of extreme rays

Abbreviations

CHO	Chinese hamster ovary
CPR	Carbon dioxide production rate
EFM	Elementary flux mode
FBA	Flux balance analysis
FVA	Flux variability analysis
iMFA	Integrated metabolic flux analysis
MFA	Metabolic flux analysis
MILP	Mixed integer linear programming
OUR	Oxygen uptake rate
RQ	Respiratory quotient

Chapter 1 Introduction

1.1 General Overview of Vaccines

The vaccine industry was estimated to be worth $61 billion in 2022 [2]. Vaccines have changed public health across the globe. Countries with high vaccination rates have found that vaccines have been responsible for the decrease in childhood diseases caused by major illnesses [3]. The World Health Organisation (WHO) estimates that 2-3 million lives are saved each year by the use of vaccines [4]. Therefore, improving production efficiency as global population increases is vital.

Vaccines utilise the ability of the human immune system to respond to, and remember, pathogenic antigens within the body [5]. A vaccine induces an immune response to protect the human body from infection and/or disease after exposure to a pathogen. This is achieved via antigens that are either derived from the pathogen or are synthetically produced to represent parts of the pathogen. Antigens can bind to a specific antibody or T-cell receptor. The antibody contacts the antigen over a broad surface and electrostatic interactions, hydrogen bonds, van der Waals forces, and hydrophobic interactions can all contribute to binding [6]. The key ingredient to most vaccines is one or more protein antigens [5, 7]. Immune response is mediated by B cells, which produce antibodies, and T cells and vaccines tend to provide protection through the induction of antibodies [7]. This learned response is then utilised by the body if the pathogen is encountered again.

1.2 Vaccine Production

How vaccines are produced varies with the vaccine type. Viral vector vaccines, grow the vaccine in the cells. They do not actually contain antigens, but instead use the body's own cells to produce them; often animal cells are used. A modified virus (vector) delivers the genetic code for the antigen to the cell. Large amounts of antigen are made by the cells when infected, which triggers an immune response within the body, creating a strong cellular response that is remembered [7, 8]. These cells are grown in bulk within bioreactors and an example of a viral vaccine is the oral polio vaccine [9].

For mRNA (messenger ribonucleic acid) vaccines, a genetic sequence containing the encoded antigen is inserted into a carrier which can replicate itself [10]. A reaction can then be triggered to synthesise the mRNA. The process allows for any sequence to be designed, produced *in vitro*, and distributed to any type of cell [11]. Two of the vaccines currently used against the SARS-CoV-2 disease are mRNA vaccines.

Inactivated vaccines require an isolated strain of the virus to be grown, often in cells, allowing it to replicate. This can be extracted and then inactivated. Inactivated vaccines are not as effective as live vaccines and often require boosters [12]. The flu vaccine is an example of an inactivated vaccine.

Regardless of the vaccine type, the time from early development to production can be 10-15 years, Figure 1.1. Clinical trials occur in three phases. The first phase gives the vaccine to a small group of people (max 100) to study side effects and immune response. The second

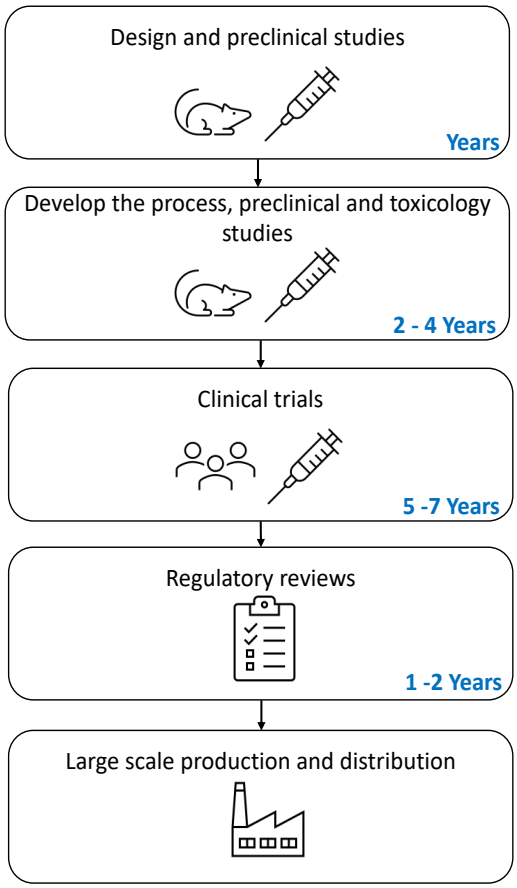

Figure 1.1 General overview of vaccine development and production timeline

phase takes hundreds of participants across a range of ages and health status like those who will receive the vaccine if successful. The final phase gives the vaccine to thousands of people to confirm the immune response, side effects and collect data on how to safely use the vaccine.

The SARS-CoV-2 pandemic led to a speed up in all steps of the production process from years to months, emphasising that if the need arose the process could be made more efficient. This, however, was only possible due to the influx in investment in the industry and a slackening in regulatory requirements [13]. The severity of the disease regarding threat to life and fast spread also necessitated a shortened vaccine development and production. Therefore, steps need to be undertaken to reduce costs and improve production efficiency when this large investment is not provided.

1.3 Vaccine Production Barriers

Vaccines, however, only work if they are administered and there are some challenges that reduce the success of this. Firstly, there are immunological challenges. Some pathogens are highly variable, such as hepatitis C, leading to genetic diversity. This makes antigen identification difficult as one antigen may not induce a response if a variation of the pathogen is encountered [5]. The RV144 vaccine for HIV was found to only offer 31% protection [14]. Secondly, the anti-vaccine movement is having a global impact. This movement does not affect the production of vaccines but does reduce the number of those wishing to be administered with one. Efforts are being made to engage with the movement to debunk myths surrounding vaccines and how they work [5].

Another issue encountered within the vaccine industry is the lack of commercial incentive for development. This occurs with diseases without global reach, such as Ebola, whereby the sale of the vaccine would not offer large profits. Reducing the operating costs in vaccine production could lead to these undesirable projects having greater chance of being tackled in industry. There is often also a lack of data available with genome networks of cells often being unknown or just recently collated. Without this information it is hard to infer how a cells' reaction route will adjust to environmental conditions. More research is needed to develop models on cells that also have extensive experimental data to allow for analysis. The final obstruction is limited

access to vaccines due to a lack of health infrastructure and poor financial resources of countries to purchase vaccines. It was found that between 2010 and 2021 global coverage for diphtheria–tetanus–pertussis-containing vaccines have only increased from 84% to 86% [15]. However, closer examination of these statistics show that some global areas had nearly 100% of children unvaccinated [5]. The Immunisation Agenda (2020-2030), of the WHO, set a clear goal to ensure that everyone is protected by full immunisation, regardless of location, age, gender, or socioeconomic barriers [16]. For this to be possible vaccines must be produced quicker and with a reduced cost; only then will most of the barriers be overcome.

To speed up production and lower costs all possible routes within the cell's metabolic network need to be known. However, at genome scale these routes are unknown due to the combinatorial explosion that occurs when trying to find them. These routes are known as elementary flux modes (EFMs) and knowledge of these EFMs allows for targeted optimisation of the process; the process in this thesis being Escherichia coli fermentation.

1.4 Escherichia coli

Escherichia coli (*E. coli*) is a foodborne pathogen that causes severe disease in humans across the globe. It is a gram-negative (thin cell wall and often highly resistant to antibiotics) [17], rod-shaped, facultative anaerobic (makes ATP via aerobic respiration) bacterium which is a prokaryotic organism [18]. A prokaryotic organism does not have a distinct nucleus with a membrane or other specialised organelles (subcellular structure that has one or more specific jobs to perform in the cell, much like an organ does in the body) [19]. *E. coli* was first described by Theodor Escherich in 1885 and since then the genome has been mapped [18, 20]. It is commonly grown in the production of recombinant proteins. Large-scale protein expression trials have shown that <50% of bacterial proteins and <15% of non-bacterial proteins are expressed by *E. coli* [21]. Recombinant proteins allow for the targeting of immune responses [22, 23].

E. coli has many advantages reported in literature and industry. Firstly, it has fast growth kinetics. In optimal environmental conditions *E. coli* doubles in 20 minutes in glucose-salt

media [24]. High cell density is easily achieved which makes it ideal for large-scale processes and finally, the complex media is readily available and inexpensive [22].

The data in this thesis was provided by GlaxoSmithKline's vaccine site in Rixensart, Belgium. Six fermentation data sets for *E. coli* are used with three antigens being produced: WT1, M72 and F4co.

1.5 Elementary Flux Modes

EFMs are non-decomposable routes through the metabolic network of a cell. A full set of EFMs describe all routes from point A to point B, like mapping routes on a road. They describe unique sets of flux carrying reactions in steady state [25]. Chapter 2 provides all the background information regarding EFMs. It is worth noting if a flux through a reaction is zero, then each contributing EFM will also have a zero flux through that reaction. Understanding the routes that a cell is using would help in improving the efficiency of production as environmental conditions could be altered to drive reactions and waste would be minimised. By increasing antigen yield and reducing waste the operating costs would reduce and vaccines could be produced at large scale quicker. Considering the length of time to get most vaccines to production and distribution, speedy production is vital.

EFMs are useful tools in setting environmental conditions for optimal production. However, finding EFMs for genome networks has not been possible. Solving EFMs for large networks causes combinatorial explosion [25]. Combinatorial explosion refers to the rapid growth of a problems complexity due to mounting constraints and bounds of the problem [26]. Across the methods created to find EFMs the maximum found thus far is just under 2 billion [27, 28, 29, 30, 31, 32, 33]. However, there is discussion regarding the need for full sets of EFMs when finding the biologically relevant ones would be more useful to process optimisation [34, 35, 36, 37]. This thesis proposes a method towards the aim of finding only active EFMs at genome scale in the future.

1.6 Aims

The aim of this thesis is to create a future-proof method towards finding active EFMs at genome scale. This aim was then reduced and is covered across 6 chapters.

Chapter 3: This chapter presents flux analysis techniques for underdetermined systems. To do this flux balance analysis and flux variability analysis is trialled. These methods are well presented in literature, and they provide a potential way of reducing the search space for EFMs. This aim was to be completed within 2 months.

Chapter 4: This chapter presents metabolic flux analysis (MFA) which can be used for exactly determinable systems. Limitations of this method needed to be identified and then an integrated form of the analysis studied, simulated, and discussed as an option to improve on the shortfalls of MFA. These techniques again offer a reduction in search space for EFMs, ensuring only active EFMs are found. This aim was to be completed in 10 months.

Chapter 5: This chapter presents a mixed integer linear programming (MILP) method to enumerate EFMs. This is performed initially on small networks and then on the *E. coli* core (in 1 hour), as presented by COBRA Toolbox [38]. Comparisons against literature presented MILP methods is also presented. This aim is met once the MILP method offers a competitive advantage to commercial tools in terms of future solving. Overall, this aim was given 18 months for completion.

Chapter 6: This chapter presents methods to improve the solve time for the MILP method presented and increase the number of EFMs found for *E. coli* core. Areas of further improvement from Chapter 5 are applied and discussed, along with methods presented throughout literature to reduce the search space for solutions. This chapter's aim is also to show the benefits of MILP in the future for computing EFMs at genome scale. This chapter is needed to reduce the solve time of MILP and once this has occurred the aim is achieved. It would take 4 months to achieve this aim.

Chapter 7: This chapter presents analysis of all the data provided by GlaxoSmithKline and assess the feasibility of determining flux and EFMs for the network. Any flux analysis that could be performed would be done and discussion regarding the effect of antigens on this data was also necessary. Additional data that should be collected to allow for further analysis would be highlighted and the need for this data emphasised. Once statement 3 in section 1.6 is met, this aim was complete, and this is done over 8 months.

1.7 Statement of Innovation

The outcomes of this thesis can be summarised into 3 main points:

1) Development of a flux analysis technique to reduce error and improve understanding of reaction rates. This technique is an integrated form of metabolic flux analysis and offers an avenue for flux analysis development in the future which has not currently been explored in literature.
2) Development of a MILP method to enumerate EFMs incorporating network compression. The MILP method allows for full genome computation in the future with computational and algorithmic advances.
3) Analysis of an *E. coli* data set provided by GlaxoSmithKline highlighted the need for further extracellular metabolite concentrations. Specific growth rates and uptake rates were calculated for all fermentations showing a reduction in these rates in the fed batch phase and a reduction in the number of feasible elementary flux modes. This information was previously unknown.

Chapter 2 Preliminaries

2.1 Introduction

Background understanding and basic definitions are key to the understanding of this thesis. This chapter will go into the basic theory behind metabolic network set up, elementary flux modes (EFMs) and a macroscopic overview of cells. It will also discuss commercial EFM solvers and network visualisation.

2.2 Metabolic Networks

The cell is equivalent to a microreactor with a high number of internal metabolites that are consumed and produced by reversible and irreversible intracellular enzymatic reactions [39]. These reactions can be mapped as a metabolic network. Metabolic networks are desirable to model due to various properties [40]:

i. There exist only small molecules within them that are near identical to one another – unlike proteins.
ii. Large quantities of the molecules within the metabolome.
iii. Interactions between the molecules have been well studied *in vitro* by organic chemists to better understand the cell environment.
iv. Metabolic models accurately capture most of their whole phenotype (observable characteristics or traits of an organism).

Metabolites in a metabolic network are categorised as intracellular or extracellular. The need for this categorisation is for the accurate modelling of the fluxes of the reactions. Extracellular metabolites are any which are outside the cell wall. However, it should be noted that there is often no distinguishing of the periplasmic space and the extracellular medium in some of the networks available [41]. In general, the term extracellular metabolites refers to the inputs and outputs of the cell [42]. Based upon this definition it can be said that intracellular metabolites are anything within the cell walls that act as intermediaries for the reactions.

Along with categorising metabolites, the reactions can also be defined in the same manner. Any reaction that crosses the cell boundary is extracellular and any reaction that occurs within the cell's walls is intracellular.

2.2.1 Stochiometric Representation

The metabolic stoichiometry can be split into the extracellular and intracellular components. As metabolism operates on a faster time scale than regulatory or cell division events, therefore allowing an assumption of quasi or pseudo steady state, a metabolite balancing equation is created [30].

$$\boldsymbol{S}_i \boldsymbol{v} = 0 \tag{2.1}$$

$$\boldsymbol{S}_e \boldsymbol{v} = \boldsymbol{v}_{\boldsymbol{m}} \tag{2.2}$$

The stochiometric matrix is represented by $\boldsymbol{S}_x$, where subscript '*i*' denotes intracellular metabolites only and '*e*' extracellular. Flux is given by $\boldsymbol{v}$, in concentration per unit of time i.e $gL^{-1}hr^{-1}$, and measurable flux by $\boldsymbol{v}_{\boldsymbol{m}}$. Equation (2.1) is the basis of all further analysis within this thesis as it ensures each metabolite is consumed in the same quantity as it is produced.

Figure 2.1 shows a simple metabolic network. The intracellular and extracellular stoichiometric matrices are,

$$\boldsymbol{S}_i = \begin{bmatrix} 1 & -1 & -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 1 & -1 \end{bmatrix} \tag{2.3}$$

$$\boldsymbol{S}_e = \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.4}$$

where, the first row in $\boldsymbol{S}_i$ is representative of metabolite 'A' and the second row, 'B'. Reactions are represented by the column entries. The numerical values being the stoichiometric coefficients of the species involved in a reaction. A negative value indicates that the metabolite is a reactant species and a positive value a product species in a particular reaction. A zero entry in the stoichiometric matrix indicates that the species does not participate in the reaction.

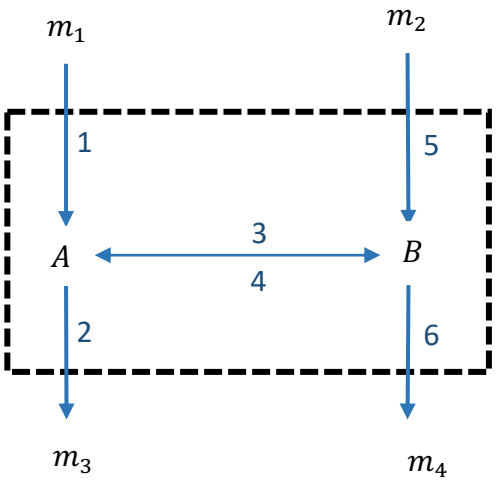

Figure 2.1 Example network to demonstrate stoichiometric representation

So, for this example there are two internal metabolites, four external metabolites and six reactions (where reactions 3 and 4 are a reversible pair). The reactions are:

$$r_1 \quad m_1 \rightarrow A$$
$$r_2 \quad A \rightarrow m_3$$
$$r_3 \quad A \rightarrow B$$
$$r_4 \quad B \rightarrow A$$
$$r_5 \quad m_2 \rightarrow B$$
$$r_6 \quad B \rightarrow m_4$$

In this thesis it is the convention to decompose a reversible reaction into two non-reversible reactions.

2.3 Elementary Flux Modes

An EFM is a minimal route through a metabolic network and are direction vectors, like the Cartesian coordinate system [25]. The word minimal implies that if the route is disrupted at any point, it can no longer continue to produce the necessary extracellular product. Therefore, EFMs allow for the production strategy to be identified for any product. No two EFMs can share the exact same reactions; they must be unique.

A complete set of EFMs will describe a metabolic network. Therefore, every steady state flux distribution can be described with a non-negative weighted superposition of EFMs [25]. EFMs

are advantageous as each flux distribution may be decomposed into fundamental units without 'cancellation' [43]. Therefore, if a flux through any reaction is zero, then each contributing EFM will have a zero-flux value for that reaction.

There are multiple uses for EFMs in industry and academia today [44]:

i. Identification of minimal conversion pathways or cycles in networks.
ii. Prediction of network properties i.e., gene essentials, blocked reactions etc.
iii. Yield optimal routes for products and biomass can be found.
iv. Identification of intervention strategies for targeted network modification.
v. Identification of module contribution in a metabolic phenotype [45].

The limitation of EFM determination is that it is a combinatorial problem which prevents large scale enumeration. To alleviate this issue several methods have been created and presented in literature. This thesis will discuss this work in detail.

EFMs can be presented as vectors or matrices. This thesis present sets of EFMs as a matrix with columns equal to the EFM and rows equal to the reactions. Table 2.1 lists the EFMs that describe the metabolic network presented in Figure 2.1 (details regarding their enumeration using traditional methods, such as efmtool, are provided later in this chapter).

Table 2.1 EFMs for example network

	$\boldsymbol{E}_1$	$\boldsymbol{E}_2$	$\boldsymbol{E}_3$	$\boldsymbol{E}_4$
r_1	1	0	1	0
r_2	1	0	0	1
r_3	0	0	1	0
r_4	0	0	0	1
r_5	0	1	0	1
r_6	0	1	1	0

There are four EFMs, $\boldsymbol{E}_1$,….,$\boldsymbol{E}_4$ (the columns in Table 2.1). Each EFM is described by the presence of a non-zero flux vector for a particular reaction; the numerical values presented in Table 2.1 for each EFM. A zero-flux vector indicating that the reaction does not participate in the EFM. Therefore $\boldsymbol{E}_1$ comprises reaction r_1 and r_2, i.e., the route

$$m_1 \rightarrow A \rightarrow m_3 \tag{2.5}$$

An alternative representation would be the unique set of reactions they represent, e.g., $E_1 = \{r_1, r_2\}$, $E_2 = \{r_5, r_6\}$, etc. The entire set of EFMs is represented in matrix notation as, $\boldsymbol{E} = [E_1 \quad \cdots \quad E_e]$.

2.3.1 Extreme Pathways

Extreme pathways (EPs) exist within metabolic networks and differ to EFMs in subtle ways. EFMS and EPs are equivalent when all reactions are irreversible and act as edges of a cone. If reversible reactions exist, then the set of EPs act as the convex basis vectors and EFMs are a superset of the EPs. Therefore, all EPs are EFMs [25, 46].

2.3.2 Flux Cones

The set of flux values contained in $\boldsymbol{v}$, that satisfy equation (2.1) exist in the null space (or Kernel) of $\boldsymbol{S}_i$. If we define,

$$\boldsymbol{S}_i\boldsymbol{K} = 0 \tag{2.6}$$

Then the matrix $\boldsymbol{K}$ consists of linearly independent column vectors that define the dependencies in the columns of $\boldsymbol{S}_i$. The null space dimension, i.e., number of column vectors, $N(\boldsymbol{S}_i)$ is found by equation (2.7).

$$N(\boldsymbol{S_i}) = N_r - rank(\boldsymbol{S_i}) \tag{2.7}$$

where N_r is the number of reactions in the network stoichiometry and $rank(\boldsymbol{S_i})$ is the rank of the stoichiometric matrix. The set of flux vectors that satisfy equations (2.1) and (2.2) can be expressed as:

$$FC = \{\boldsymbol{v} \in \mathbb{R}^{\boldsymbol{n}} | \boldsymbol{S_i v} = 0, \boldsymbol{v_{irr}} \geq 0\} \tag{2.8}$$

where $\boldsymbol{v_{irr}}$ are irreversible fluxes. Equation (2.8) is a subset of the null space. It is the intersection of the null space with the nonnegative half spaces corresponding to irreversible reactions [44]. Geometrical speaking, this is a convex polyhedral cone. In this work, defined as the flux cone (FC). A general polyhedral cone is defined as equation (2.9).

$$C = \{\boldsymbol{x} \in \mathbb{R}^{\boldsymbol{n}} | \boldsymbol{Ax} \geq 0\} \tag{2.9}$$

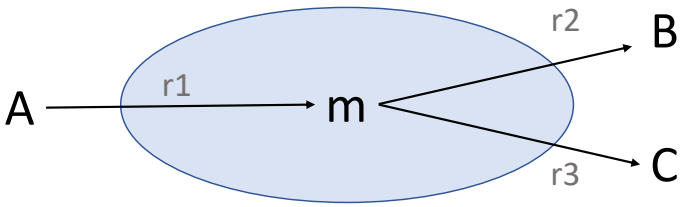

Figure 2.2 A small reaction network with three extracellular metabolites (A, B, C) and one intracellular (m).

Figure 2.2 illustrates a small network example. The stoichiometric matrix for this network is,

$$\boldsymbol{S} = \begin{bmatrix} -1 & 0 & 0 \\ 1 & -1 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.10}$$

For this example, the intracellular stoichiometric matrix is,

$$\boldsymbol{S}_i = [1 \quad -1 \quad -1] \tag{2.11}$$

The rank of this matrix is one and as there are three reactions the null space dimension is 2. Therefore, any two linearly independent vectors will form the basis of the null space. For example,

$$\boldsymbol{K} = \begin{bmatrix} 1 & 0 \\ 2 & -1 \\ -1 & 1 \end{bmatrix} \tag{2.12}$$

Note that these basis vectors do not have to meet the irreversibility constraint as set in equation (2.8) [44].

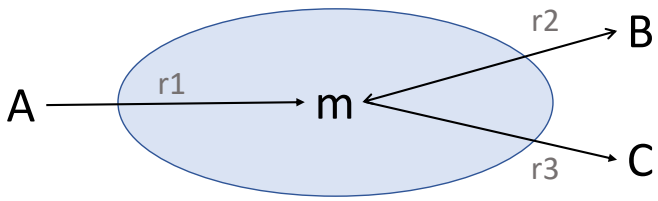

Figure 2.3 A small reaction network with three extracellular metabolites (A, B, C),one intracellular (m) and a reversible reaction

Figure 2.3 is the same network as presented in Figure 2.2, however reaction 2 is now deemed reversible. There exists three EFMs for this network, $\boldsymbol{E_1} = [1\ 1\ 0]^T$ and $\boldsymbol{E_2} = [1\ 0\ 1]^T$ and $\boldsymbol{E_3} = [0\ -1\ 1]^T$. Figure 2.4 illustrates the flux cone which is bounded on one side by two extreme vectors $\mathbf{e_1}$ and $\mathbf{e_2}$. These extreme vectors are also EFMs, along with $\mathbf{e_3}$. Often in metabolic analysis it is common to only be interested in minimal routes within the network [44, 47]. This creates a basis for characterising the flux cone. However, if this is done often EFMs that exists within the space and are not the bounds are missed, so $\mathbf{e_3}$ would not be found. It therefore is imperative that any method to find EFMs is inclusive of these vectors.

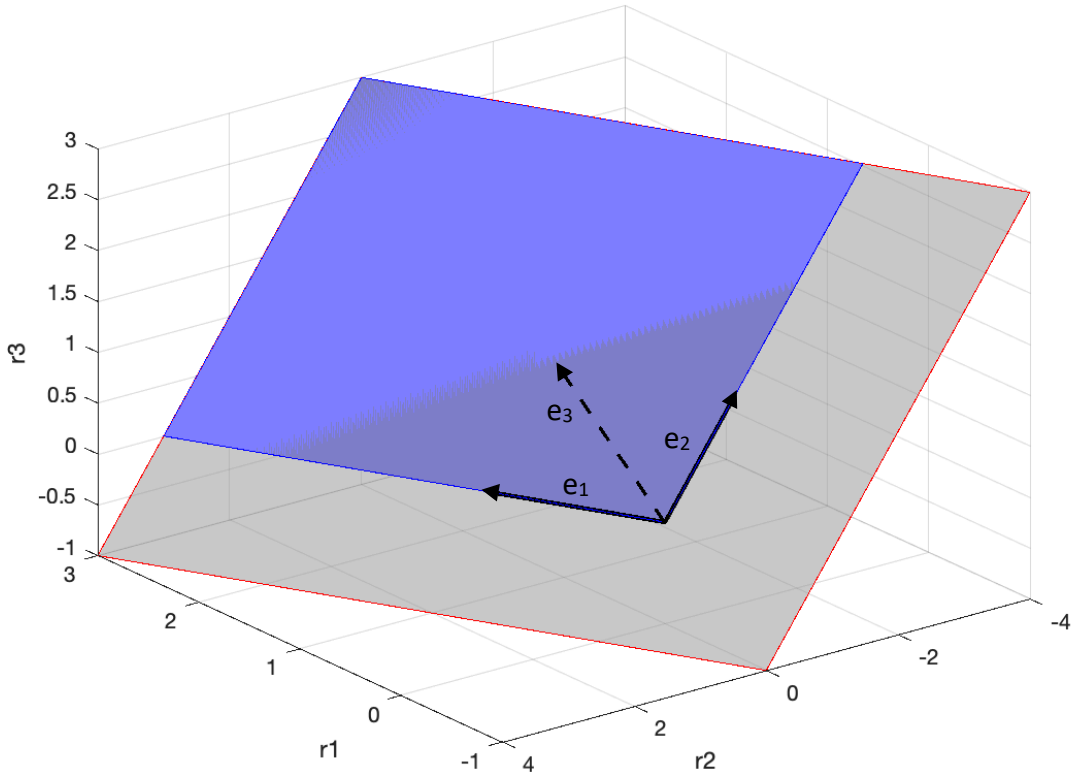

Figure 2.4 The null space of example network in figure (2.1). The red lines indicate that the null space is unbounded (in all directions). The black arrows are extreme vectors (e1, e2) and EFMs (all e).

2.3.3 Weighting of Elementary Flux Modes

The metabolic network of a cell can be conceived as the weighted sum of EFMs [39]. Flux distribution across a network can be decomposed into the EFMs by equation (2.13),

$$\boldsymbol{v} = \sum_{j=1}^{\boldsymbol{E}} w_j \boldsymbol{E}_j \tag{2.13}$$

where w_j is a weighting associated with the EFM vector, $\boldsymbol{E}$. The weightings are commonly known as elementary mode weightings [48]. The elementary mode weightings are expressed in units of flux. To represent the fractional usage of each EFM instead, an α-weighting can be used. An α-weighting is a weighting that is normalised by dividing the fluxes by the limiting maximum flux of each EFM [49].

In this thesis an EFM can only be irreversible, therefore the weighting will only ever be positive,

$$w_j \geq 0 \tag{2.14}$$

As the number of EFMs in a system increase equations (2.13) and (2.14) do not provide a unique solution. Instead they offer a continuous convex space of feasible solutions [48]. To allow for unique solutions to be obtained equation (2.8) is necessary to apply minimise a constraint [48], equation (2.15).

$$min \sum_{j=1}^{E} w_j{}^2 \tag{2.15}$$

For an example, the network in Figure 2.2 may have a flux distribution of,

$$\boldsymbol{v} = \begin{bmatrix} 6 \\ 4 \\ 2 \end{bmatrix} \tag{2.16}$$

This flux vector indicates that all 3 reactions are in operation. Utilising the two EFMs that exist for this network,

$$\boldsymbol{E} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{2.17}$$

The weightings for each EFM required to generate the flux as close to the specified distribution are,

$$\boldsymbol{w} = \begin{bmatrix} 4 \\ 2 \end{bmatrix} \tag{2.18}$$

Therefore, both EFMs are necessary to get the specified flux distribution. The residual flux from for each reaction using the specified weightings in this example is,

$$\boldsymbol{v_{residual}} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{2.19}$$

2.4 Elementary Flux Mode Solvers

There are many tools available to find EFMs all with varying efficiency, operating system capability and limitations. The three most used throughout literature are efmtool [29, 31], METATOOL 5.0 [50, 51] and FluxModeCalculator [33]. All tools are designed to operate on Windows, with efmtool also working on Mac and Linux. These tools vary in memory usage so their operating times on different computer types will vary. These tools will be compared alongside the mixed integer linear programming method proposed within this thesis.

2.4.1 METATOOL 5.0

METATOOL uses a variant on the double descriptive method [50]. Like efmtool, it can be used in MATLAB allowing basic changes to be made by the user. The notable difference with this tool is that the test for checking the independence of EFM candidates is performed algebraically, via a rank test [50, 52]. It also computes structural invariants such as conservation relationships and enzyme subsets [50, 51].

This algebraic test is notably quicker than a combinatorial test [27].The combinatorial test compares each EFM candidate to all other modes, leading to the test growing as the number of EFMs grow. Rank testing avoids this as it is independent of the other modes and the complexity is limited to the number of reactions used by the EFM. The rank test will be discussed in Chapter 4, as it is used for the development of the MILP.

METATOOL was previously the fastest solver available for efm enumeration, however, it has now been reported to be outperformed by efmtool [31].

2.4.2 efmtool

This tool uses the double descriptive method (described in detail in section 2.6) as it's basis and is coded into the MATLAB environment. EFM-candidates are found via pairwise combinations and must be verified to be an EFM afterwards. The verification is the major bottleneck of the double descriptive method [25]. Over the years there has been multiple improvements such as exploiting multi-core CPUs to allow for greater storage capabilities and bit set trees to speed up computation [31]. The tool benefits from easy integration with the

COBRA toolbox [38]. COBRA provides the *E. coli* core network and can perform flux analysis with specific data input. This toolbox is used in Chapter 7 of this thesis. Overall efmtool has been shown to be the best performing EFM solver operating on a single processor machine [53, 54] .

2.4.3 FluxModeCalculator

The double descriptive method is used for FluxModeCalculator with improvements from literature being utilised, such as network compression, solved inequalities stored as bit patterns and the use of bit pattern trees for combinatorial testing [33]. It is stated to require 4GB of free memory to calculate 271 million EFMs in *E. coli* in 17 hours [33]. This is partially due to the solver saving immediate solutions on the hard drive to reduce the memory size required to completed one iteration. Moreover, it checks at the end of each iteration whether the resulting flux modes still fit in the memory. If not, the system terminates providing the intermediate solution. Along with this it also uses a demand based network subdivision strategy to automatically subdivide the network for the remaining constraints and calculated the flux modes for the corresponding subnetworks [28].

2.5 Double Descriptive Method

All the above tools use the double descriptive method as their basis. Originally the double descriptive method was proposed by Motzkin *et al* for the determination of a numerical value and of all solutions in a game with a finite number of strategies, and of general finite systems with linear inequalities and corresponding maximisation problems [55]. It is now widely used as a basis for calculating EFMs of a metabolic network. It is able to find all extreme rays of a general polyhedral cone [56]. The double descriptive method works by finding double descriptive pairs – a pair of matrices that contain two different descriptions of the same object. This is useful for the solving of EFMs as an EFM always corresponds to an extreme ray, Figure 2.5 [57]. The final cone, like the one in Figure 2.5, is formed through the application of constraints. Each constraint is represented by a half space, which intersects the original cone until all constraints are accounted for. Figure 2.6 provides a visual example of this, showing a new extreme ray, h, being created from adjacent rays ‘a’ and ‘g. Any descendants from non-adjacent rays, like ‘i’ from ‘c’ and ‘g’, are not extreme rays.

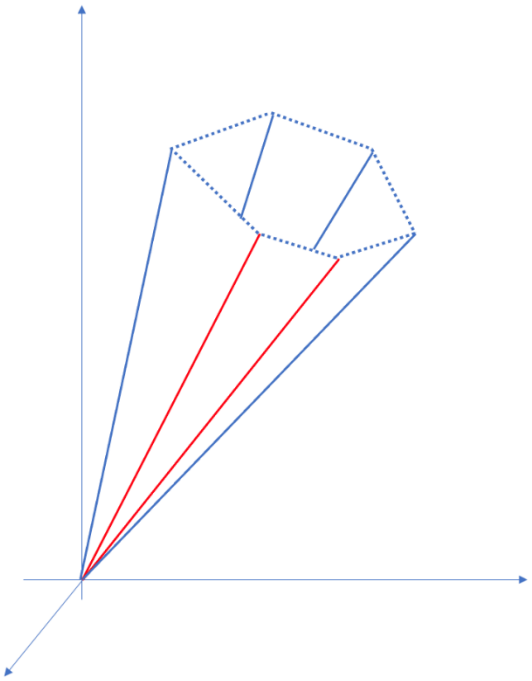

Figure 2.5 Pointed polyhedral cone. Dashed lines represent unbounded areas. Red lines highlight a pair of adjacent extreme rays

Therefore, every intersection with half space 'H' removes extreme rays, creating a smaller cone each iteration. New extreme rays are created by adjacent extreme ray pairs through Gaussian elimination [31]. The new ray lies in the hyperplane, separating retained from removed rays.

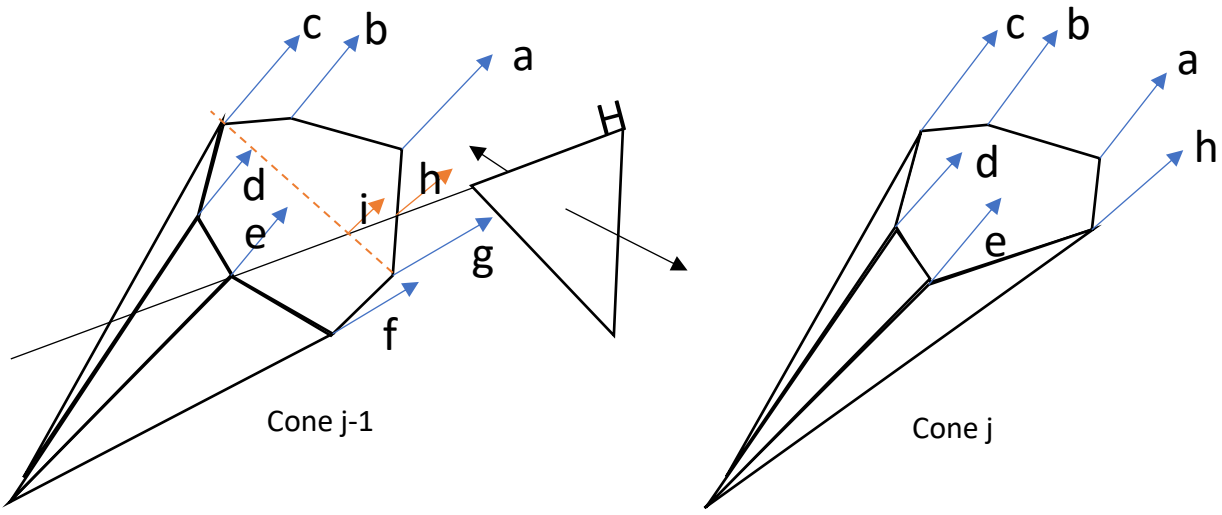

Figure 2.6 Graphical representation of the double description method finding EFMs

Wagner *et al* proposed using a unique form of the null space of the stochiometric matrix as the initial cone, equation (2.20) [58]. A kernel ($\boldsymbol{K_r}$) of a matrix is the same as the null space of a matrix.

$$\boldsymbol{K_r} = [\boldsymbol{I}; \boldsymbol{K_r}^*]^T \tag{2.20}$$

$\boldsymbol{I}$ is the identity matrix and this accounts for the irreversibility of fluxes. The steady state constraint equation (2.1) is met by $\boldsymbol{K_r}$. After applying these constraints, a reduced cone, 'j' is found, like in Figure 2.6. Extreme rays can be found iteratively and are grouped into one of three categories: positive, negative or zero flux. To create the next, reduced cone (j+1), extreme rays with positive and zero flux are kept, and new ones generated from ray pairs. To create new pairs the flux value at position j+1 is cancelled out, to be 0, via the use of Gaussian elimination.

Due to the new ray being a combination of old rays, non-negative and Definition 1 being true, the new ray is a ray of a cone.

Definition 1[31]

A set P of points in $\boldsymbol{R}^d$ *is convex if the line segment between two points in P lies in P. A set P is referred to as a cone if for every* $\boldsymbol{x} \in P$*, it's non-negative multiple lies in P. It is also a convex cone if:*

$$\lambda_1 \boldsymbol{x}_1 + \lambda_2 \boldsymbol{x}_2 \in P$$
$$\lambda_1, \lambda_2 \geq 0$$
$$\boldsymbol{x}_1, \boldsymbol{x}_2 \in P$$

Adjacency tests are needed to make sure that the ray found is an extreme ray and thus an EFM. To ensure this equation (2.21) must be upheld, where Z is the set of inequality indices satisfied by the extreme ray with equality [57]. Therefore, $Z(\boldsymbol{e})$ and $Z(\boldsymbol{e}')$ contain at least one element that is the same and $Z(\boldsymbol{e}'')$ is fully contained, a subset of, or equal to $Z(\boldsymbol{e}')$.

$$Z(\boldsymbol{e}) \cap Z(\boldsymbol{e}') \subseteq Z(\boldsymbol{e}'') \tag{2.21}$$

Most solvers use a variation on the double descriptive method to find EFMs. They often only require user input of stoichiometry and reaction reversibility making them much more efficient than performing the double descriptive method by hand. It is worth noting that any method that uses the double descriptive method as the base for solving EFMs will be sensitive to constraint ordering [59]. Constraint ordering refers to the ordering of matrix rows when solving with the double descriptive method. Commonly the null space approach is used, equation (2.6) to generate the kernel matrix in row-echelon form to act as the first extreme ray matrix. The identity matrix must be preserved; however, all remaining rows must be organised to maximise efficiency [31].

2.5.1 Double Descriptive Method Example

For the hypothetical metabolic network postulated in for Figure 2.7, the intracellular stochiometric matrix ($\boldsymbol{S}_i$), irreversible reaction only matrix ($\boldsymbol{S}_{\boldsymbol{irr}}$) and reversible reaction only matrix ($\boldsymbol{S}_{\boldsymbol{rev}}$) are as follows:

$$\boldsymbol{S}_i = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & -1 \\ 0 & 1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 1 & 0 & -1 & 1 \end{bmatrix} \tag{2.23}$$

$$\boldsymbol{S}_{\boldsymbol{irr}} = \begin{bmatrix} 1 & 0 & -1 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \tag{2.24}$$

$$\boldsymbol{S}_{\boldsymbol{rev}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.25}$$

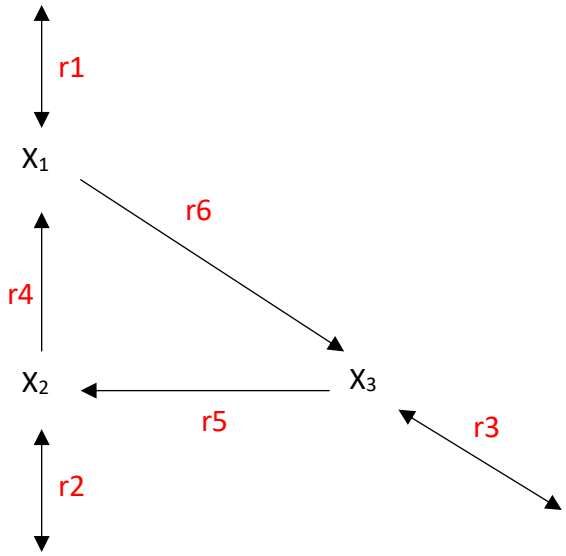

Figure 2.7 Hypothetical metabolic network example

A tableau ($\boldsymbol{T}^{(m)}$) can be formed and can be solved up to a max of '*m*' iterations, equation (2.26) [60]. In the tableau $\boldsymbol{B}^{(m)}$ represents the reversible reaction section and $\boldsymbol{F}^{(m)}$ irreversible reaction section. The aim of the iterations is to get the stochiometric matrices to be filled with zero elements in '*m*' steps. This in turn creates a set of EFMs. The reactions for larger networks may also be grouped together based upon their reliance on one another to reduce the size of the tableau [61]. Equation (2.26) is better interpreted by relating it to the steady state constraint, equation (2.27). By forcing the stochiometric matrices to be 0 ensures only a set of EFMs will be found, equation (2.28)

$$\boldsymbol{T}^{(\mathbf{0})} = \begin{pmatrix} \boldsymbol{B}^{(\mathbf{0})} \\ \boldsymbol{F}^{(\mathbf{0})} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I} & \mathbf{0} & \boldsymbol{S}_{\boldsymbol{rev}}^{\boldsymbol{T}} \\ \mathbf{0} & \boldsymbol{I} & \boldsymbol{S}_{\boldsymbol{irr}}^{\boldsymbol{T}} \end{pmatrix} \tag{2.26}$$

$$\boldsymbol{S}_i \boldsymbol{v} = 0 \tag{2.27}$$

$$\mathbf{0} = \begin{pmatrix} \boldsymbol{B}^{(\mathbf{0})} \\ \boldsymbol{F}^{(\mathbf{0})} \end{pmatrix} = \begin{pmatrix} \boldsymbol{v}_{\boldsymbol{rev}} & \boldsymbol{S}_{\boldsymbol{rev}}^{\boldsymbol{T}} \\ \boldsymbol{v}_{\boldsymbol{irr}} & \boldsymbol{S}_{\boldsymbol{irr}}^{\boldsymbol{T}} \end{pmatrix} \tag{2.28}$$

To combine rows, equations (2.29) and (2.30) must be upheld. These equations are for elements in a column number, that is equivalent to the iteration number, 'm'. However, the whole row must undergo the same operation as the elements concerned. A scalar can be used on any row to ensure these equations are met.

$$f_{ij} \pm b_{ij} = 0 \qquad (2.29)$$
$$f_{ij} \neq 0 \; AND \; b_{ij} \neq 0$$

$$f_{ij} + f_{*ij} = 0 \qquad (2.30)$$
$$f_{ij} < 0 \; AND \; f_{*ij} > 0$$

Where:

i = row number

j = column number, which is equivalent to the iteration you are currently undergoing

b = element in row vector of $\mathbf{B}$

f = element in row vector of $\mathbf{F}$

$*$ i = a differing row number to i

To ensure only EFMs are found, equation (2.21) must be used; no two rows can be combined if there exists another row in the same space. Equation (2.29) combines any row of $\boldsymbol{B}$, with a value not equal to 0 in the current column, with all rows in $\mathbf{F}$, which must also not be equal to 0. Equation (2.30) combines all rows in $\mathbf{F}$ in pairs if they have opposite signs in the current column.

The first tableau, $\boldsymbol{T}^{(0)}$, is set up as follows (the blue dotted lines are to aid in the identification of the different sections, blue text provides row number):

$$\boldsymbol{T}^{(0)} = \left[\begin{array}{cccccc:ccc} 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ \hdashline 0 & 0 & 0 & 1 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 1 \end{array}\right] \begin{array}{c} 1 \\ 2 \\ 3 \\ \hdashline 4 \\ 5 \\ 6 \end{array}$$

To create the next tableau, $\boldsymbol{T}^{(1)}$, the first column of the right-hand-side quadrants is examined. Any row that has a zero element in this column does not need to be combined with any other row. Therefore, by examining $\boldsymbol{T}^{(0)}$it can be said that only rows 1,4 and 6 must be used further.

To create the next tableau the following combinations, based upon equations (2.21), (2.29) and (2.30), are performed:

i. Row 1 and row 4. The right-hand-side has + 1 values in the first column, so these vectors must be subtracted from one another, equation (2.29).
ii. Row 1 and row 6. The right-hand-side has differing signs so the two rows can just be summed together.

A similar process is adapted for section $\boldsymbol{F}$. Any rows in $\mathbf{F}$ with opposing signs in the current column should be combined. Thus, row 4 and 6 can be combined via summation due to opposing signs in the right-hand-side initial column and but also due to their independence compared to all other row vectors.

Any row vector that is combined with another is removed form the next tableau as they are accounted for via the combination vector created. It is important to remember that section $\boldsymbol{B}$ accounts for reversible reactions, and section $\mathbf{F}$ the irreversible ones. Therefore, when creating the next tableau, the new vector must be placed in the correct section. Any vector combination that includes a vector from $\mathbf{F}$ must remain in $\mathbf{F}$. This is due to the vector no longer being fully reversible.

The new tableau to examine is as follows (red text shows how the vectors have been formed based upon previous row numbers):

$$\boldsymbol{T}^{(1)} = \left[\begin{array}{ccccccc|cc|cc} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 2 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 3 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & -1 & 3 & 5 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & -1 & 1 & 4 & 4+6 \\ -1 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 5 & 1-4 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 6 & 1+6 \end{array}\right]$$

This process is continued until the right-hand-side is a zero matrix, $\boldsymbol{T}^{(3)}$:

$$\boldsymbol{T}^{(3)} = \boldsymbol{F}^{(3)} = \left[\begin{array}{cccccc|ccc} -1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 1 & 0 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \end{array}\right] = (\boldsymbol{f}^k \vdots \mathbf{0})$$

Therefore, there are 7 EFMs for this network, all classed as irreversible (all exist in section $\mathbf{F}$). Negative stoichiometry gives directionality to the reversible reactions. Translation of these results into the form used in this thesis is given in Table 2.2.

Table 2.2 EFMs for double descriptive method

	$\boldsymbol{E}_1$	$\boldsymbol{E}_2$	$\boldsymbol{E}_3$	$\boldsymbol{E}_4$	$\boldsymbol{E}_5$	$\boldsymbol{E}_6$	$\boldsymbol{E}_7$
$r_1: \rightarrow X_1$	1	0	1	0	0	0	1
$r_2: X_1 \rightarrow$	0	0	0	0	1	0	0
$r_3: X_2 \rightarrow X_1$	0	0	0	1	1	1	0
$r_4: X_3 \rightarrow X_2$	0	1	0	0	1	1	1
$r_5: X_1 \rightarrow X_3$	1	0	1	1	0	1	1
$r_6: X_2 \rightarrow$	0	1	0	0	0	0	1
$r_7: \rightarrow X_2$	0	0	0	1	0	0	0
$r_8: X_3 \rightarrow$	1	0	1	1	0	0	0
$r_9: \rightarrow X_3$	0	1	0	0	1	0	0

2.6 Network Visualisation

In chemical work molecular graphs are used to better understand the structure of a molecule. The graph is made up of nodes and edges, where the nodes are the atoms, and the edges are the bonds between the atoms. This is then applied to chemical reaction networks; a reaction network is effectively a flow network. In this network the elements are the nodes, and the reactions are the edges. Therefore, it can be easy to apply this to metabolic networks. However, metabolites can participate in more than one reaction and reactions can have more than one substrate and more than one product.

Directed graphs allow for both metabolic network and EFM representation. Visualisation of EFMs on publicly available tools is limited. EFM data is produced in a table or matrix format, this data must then be exported to online tools, such as Escher, to visualise the EFMs on the network [38]. With large data sets this data can be meaningless, however, an integrated visualisation tool would make it more accessible.

These graphs have been applied to analyse the metabolism of cells in greater detail. For example, degree distributions and centrality measures can highlight network connectivity [62, 63, 64]. Deletion of any nodes or edges can be representative of environmental changes or therapeutic drugs which target specific metabolite enzymes [65, 66]. Often these graphs are preferred over flux balance analysis, which maximises an objective function, as the graph relies solely on the metabolic stoichiometry [64].

Metabolic networks can be translated into graphs by one of three ways: 1) reaction centric, where reactions are the nodes and metabolites the edges, equation (2.9) [67], 2) metabolite centric, where metabolites are the nodes and reactions are the edges, equation (2.10) [68] or 3) bipartite, where both reactions and metabolites act as nodes [64].

To construct a network graph, consider a stoichiometric matrix that is written in terms of the reactant and product stoichiometric numbers,

$$\boldsymbol{S} = [\boldsymbol{S}_p - \boldsymbol{S}_r] \tag{2.31}$$

So, for example, for a metabolic network described by the stoichiometry,

$$\boldsymbol{S} = \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 1 & -1 & -2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 & -1 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.32}$$

There are seven reactions, $r_1, \ldots . r_7$ and ten metabolites with the naming A, G, F, B, C, E, D, H, I, P. The external metabolites are A, G, F (substrates) and H, I, P (products). The intracellular metabolites are, B, C, D, E. The matrices of the stoichiometric numbers of the reactant and product species are given by,

$$\boldsymbol{S}_r = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{2.33}$$

$$\boldsymbol{S}_p = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.34}$$

Using a binary representation of $\boldsymbol{S}_r$ and $\boldsymbol{S}_p$ (where all the nonzero stoichiometric numbers are indicated by ones, gives two matrices denoted $\widehat{\boldsymbol{S}}_r$ and $\widehat{\boldsymbol{S}}_p$) the reaction centred, $\boldsymbol{A}_R$, and metabolite centred, $\boldsymbol{A}_M$, adjacency matrix is defined as,

$$\boldsymbol{A}_{R_{(N_R,N_R)}} = \widehat{\boldsymbol{S}}_p^T \widehat{\boldsymbol{S}}_r \tag{2.35}$$

$$\boldsymbol{A}_{M_{(N_M,N_M)}} = \widehat{\boldsymbol{S}}_r \boldsymbol{S}_p^T \tag{2.36}$$

For the stoichiometric matrix, equation (2.32), this gives,

$$\boldsymbol{A}_R = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{2.37}$$

$$\boldsymbol{A}_M = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{2.38}$$

The adjacency matrices can be used to draw the directed graphs that connect nodes of the graph, where the non-zero terms in the adjacency matrix define the connections between the nodes (edges of the graph). For example, for the reaction centric adjacency matrix, with nodes, $r_1, \ldots., r_7$ the following edges are defined (which can be verified by inspection of the rows of the adjacency matrix),

$$(r_1, r_2), (r_1, r_3), (r_3, r_6), (r_3, r_7), (r_4, r_2), (r_5, r_6)$$

Similarly, for the metabolite centric adjacency matrix, with nodes A, G, F, B, C, E, D, H, I, P the following edges are defined,

$$(A, B), (G, E), (F, D), (B, C), (B, H), (C, I), (C, P), (E, I), (D, H)$$

The reaction centric digraph is shown in Figure 2.8 and the metabolite centric in Figure 2.9.

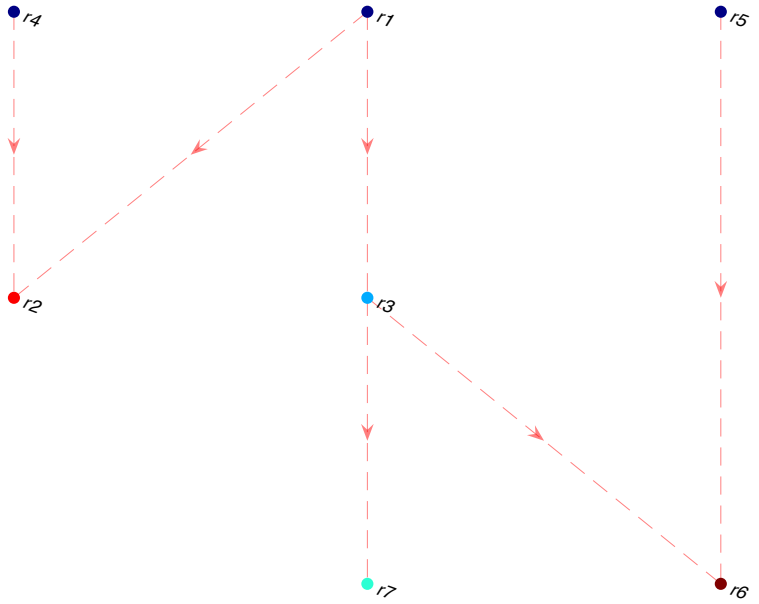

Figure 2.8 Reaction centric digraph

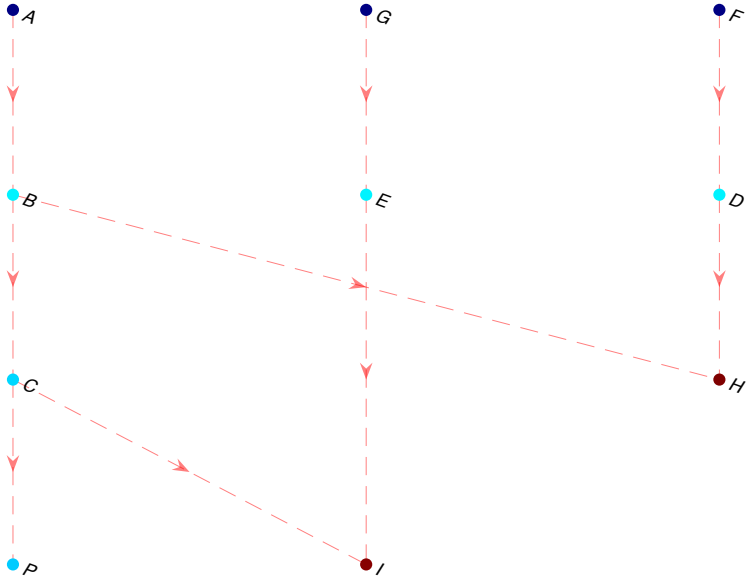

Figure 2.9 Metabolite centric digraph

An alternative to the reaction or metabolite centric graph is the use of a bipartite graph that includes both reactions and metabolites (also referred to as a hyper-graph). The adjacency matrix of the bipartite graph is given by,

$$A_{M,R} = \begin{bmatrix} \mathbf{0} & \widehat{S}_r \\ \widehat{S}_p^T & \mathbf{0} \end{bmatrix} \tag{2.39}$$

This has nodes ordered by metabolites and then reactions, and for the example being considered the adjacency matrix is a 17 x 17 matrix (10 metabolites and 7 reactions). Figure 2.10 shows the bipartite graph created for this example.

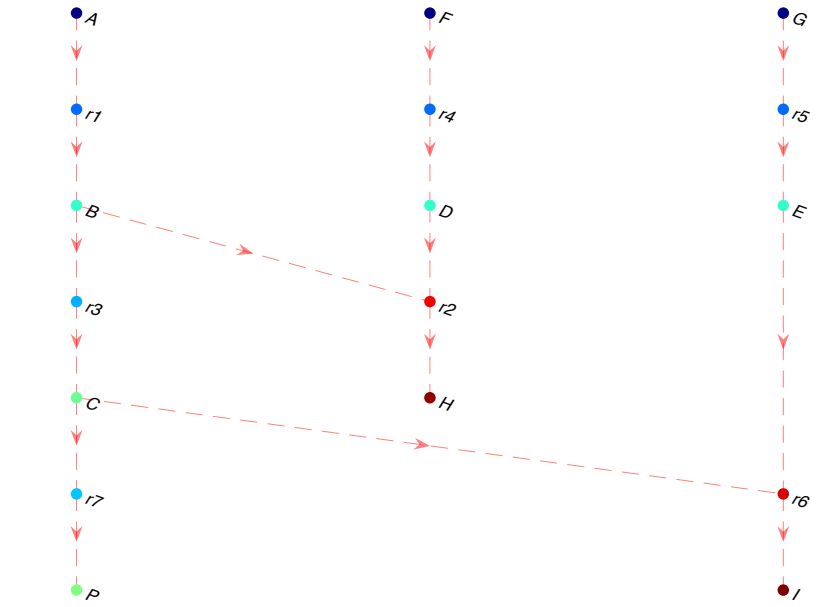

Figure 2.10 Bipartite digraph

2.7 Dead-end Metabolites

A dead-end metabolite (DEM) is an intracellular metabolite that is produced by known reactions in the network but has no reactions consuming it or is consumed by known reactions in the network but has no known reactions producing it. In both these cases there are no identified transporter reactions [69]. DEMs can occur due to a lack of knowledge of network structure where further experimental research is required. Figure 2.11 highlights three dead-end metabolites, H, I and K. EFMs will not be generated containing the reactions that lead to or from these metabolites as any EFM should go from extracellular substrates to products. Therefore, removal of these reactions from any stochiometric matrix will reduce the search space for EFMs.

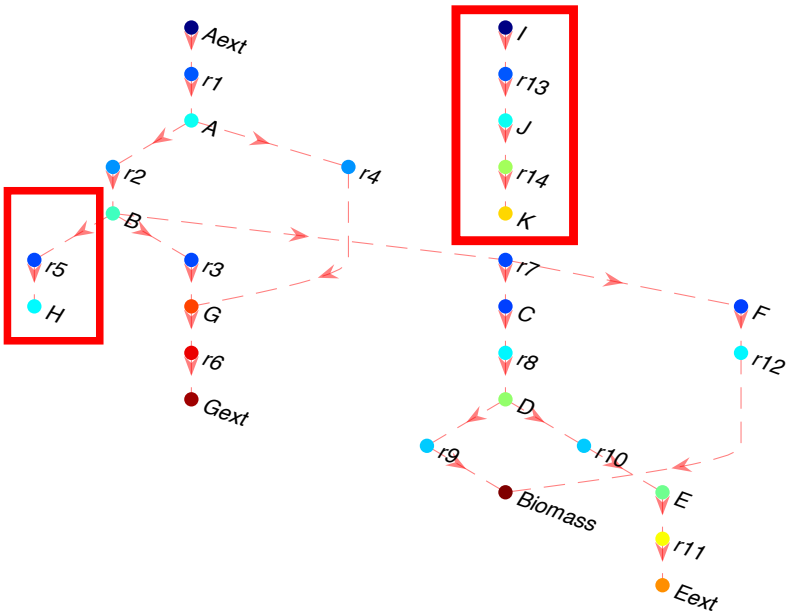

Figure 2.11 Bipartite graph network with dead-end sections highlighted in red

2.8 Macroscopic Networks

The overall stochiometric matrix for the macroscopic network of a cell is found via equation (2.40). If the rank of $\boldsymbol{S_K}$ is less than the number of unknowns, the system is undeterminable.

$$\boldsymbol{S_K} = \begin{bmatrix} \boldsymbol{S_S} \\ \boldsymbol{S_P} \end{bmatrix} \times \boldsymbol{E} \tag{2.40}$$

$\boldsymbol{S_K}$ is the overall (macroscopic) stoichiometry, $\boldsymbol{S_S}$ is the stoichiometry of the substrates only and $\boldsymbol{S_P}$ is the stoichiometry of the products only. The number of macro reactions for any given network is equivalent to the number of EFMs. Reaction rate equations can be designed specifically for macro reactions. The maximum specific uptake/excretion rates for extracellular metabolites, $\boldsymbol{a_i}$, are therefore related to $\boldsymbol{S_K}$ via equation (2.41).

$$\boldsymbol{v_m} = \boldsymbol{S_K} \times \boldsymbol{a_i} \tag{2.41}$$

Accurate determination of EFMs enables a macro view of a cell to be obtained. This in turn can be used to create dynamic models of the cell's lifetime. This is discussed in further detail in Chapter 3.

2.9 Linear Programming

A general, finite dimensional, optimisation program can be defined as,

$$\begin{aligned} & min_x\, f(x) \\ s.t.\, & g_i(x) \leq 0, i = 1, \dots, m \\ & h_j(x) = 0, j = 1, \dots, m \\ & x \in D \end{aligned} \tag{2.42}$$

where, $D \subseteq \mathbb{R}^n$, $g_i : D \rightarrow \mathbb{R}$ for $i = 1, \dots, m$ and $f : D \rightarrow \mathbb{R}$. $f(x)$ is the objective function, $g_i(x) \leq 0$ a set of inequality constraints and $h_j(x) = 0$ a set of equality constraints. Rather than minimising equation (2.42) you could also maximise. Rather than a general optimisation problem, this work considers linear programs (LPs), therefore both the objective function and constraints are linear. This is defined in matrix and vector forms as,

$$\begin{aligned} & min_x\, c^T \boldsymbol{x} \\ s.t.\, & \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \\ & \boldsymbol{A_{eq}}\boldsymbol{x} = \boldsymbol{b_{eq}} \\ & \boldsymbol{x} \in \mathbb{R}^n \end{aligned} \tag{2.43}$$

where, $c \in \mathbb{R}^n$, $\boldsymbol{A}$ and $\boldsymbol{A_{eq}} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{b}$ and $\boldsymbol{b_{eq}} \in \mathbb{R}^m$. All LPs can be written in the form as presented in equation (2.43). Any vector $\boldsymbol{x}^{\#}$ is a feasible solution if it fulfils all the constraints set by the LP. The solution space contains all feasible $\boldsymbol{x}^{\#}$s. However, an $\boldsymbol{x}^{\#}$ is only an optimal feasible solution, $\boldsymbol{x}$, if $c^T \boldsymbol{x}^{\#} \leq c^T \boldsymbol{x}$ for all $\boldsymbol{x} \in \mathbb{R}^n$ within the set constraints. There is no requirement for a feasible solution to exist for a LP and if a feasible solution does exist, there does not need to be an optimal solution [70]. Optimal solutions are also not required to be unique.

LP constraints are equivalent to a half space in $\mathbb{R}^n$ which reduces the solution space. The intersections of these half spaces define a polyhedron, Figure 2.12.

The simplex method, proposed by Dantzig in 1948, is the most used way to solve LPs [71]. It moves from one vertex to an adjacent vertex, whilst improving the objective value, until reaching an optimal solution [72]. If the problem is unbounded or infeasible the algorithm will stop with no solution produced.

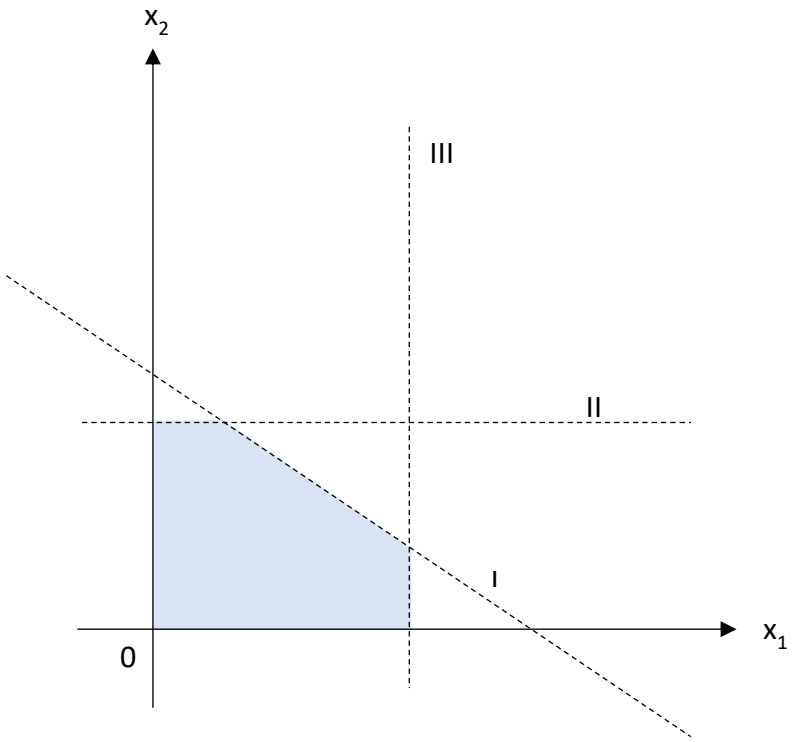

Figure 2.12 Constraint (I, II and III) application to define a polyhedron

2.10 Mixed Integer Linear Programming

If any of the variables in a LP are integers, such as binary variables, mixed integer linear programming (MILP) must be used. Binary variables can be used to define on/off constraints. MILP has been shown to be NP-hard (nondeterministic polynomial time) [73]. Equation (2.42) is still upheld but a subset of the decision variables, x, are defined as binary variables, i.e., $\{0,1\}$.

The most common method of solving MILPs is the branch and bound method [74]. The solver iteratively goes down various paths, or branches, reaching a feasible solution within the constraints provided, Figure 2.13. The branch and bounding method is a solution approach that is based upon the principle that the total set of solutions for a problem can be reduced to a smaller subset of solutions. These subsets can then be evaluated until the best solution is found. The success of branch and bounding lies in guiding the initial search [75]. Once a

solution is found it saves it and repeats the process. For larger networks, this branching and bounding process is time consuming and can cause combinatorial explosion.

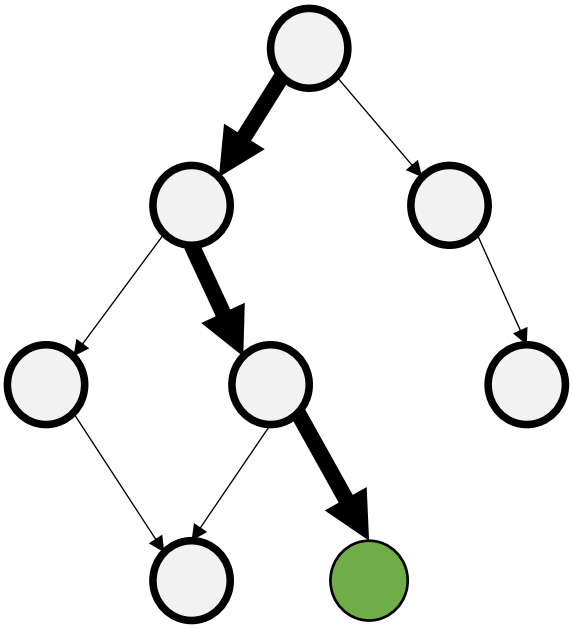

Figure 2.13 Branch and bounding of MILP. The green circle highlights a feasible solution

2.11 Summary

EFMs give an overall map of all routes possible in a metabolic network. These EFMs can be found within a convex polyhedral cone. Most commercial solvers use the double descriptive method to find these EFMs. However, it still has not been possible to find genome scale network EFMs due to the combinatorial explosion that occurs whilst solving. Metabolic networks can be drawn either reaction or metabolite centrically or as a bipartite graph. These allow users to observe how reactions are connected to metabolites through the network and can even highlight dead-end metabolites. Removal of dead-end metabolites reduces the network size, and this will help in the detection of EFMs, particularly with linear programming techniques.

Chapter 3 Flux Balance and Flux Variability Analysis

3.1 Introduction

Flux analysis is used in metabolic engineering to maximise product yield, find the range of feasible fluxes and to estimate unknown intracellular flux values. There are two well documented methods that will be discussed in this chapter:

1. Flux balance analysis (FBA) – optimisation of one metabolite through a reaction with other flux values estimated around this.
2. Flux variability analysis (FVA) – range of feasible flux values are found for an underdetermined system.

A system is underdetermined if,

$$\mathrm{rank}(\boldsymbol{S_N}) < \mathrm{u} \tag{3.1}$$

where $\boldsymbol{S_N}$ is the stoichiometry consisting of the reactions with unknown fluxes and u is the number of unknowns [76]. Underdetermined systems do not allow for unique flux solutions to be obtained [76].

3.2 Flux Balance Analysis

FBA has a wide range of uses due to its simplicity, from gap filling networks to genome-scale synthetic biology [77]. There is also often no need to code FBA from scratch as tools such as COBRA provide a reliable system for doing so [38]. However, the method is easily performed within the framework of linear programming introduced in Chapter 2. FBA is performed using the pseudo steady state constraint, equation (3.2), where $\boldsymbol{S}_i$ is the intracellular metabolic stoichiometry, $\boldsymbol{S}_e$ is the extracellular metabolic stoichiometry and $\boldsymbol{v}$ is the flux vector for all reactions in the metabolic network. Three constraints are required:

i) the flux is irreversible, equation (3.2).
ii) the flux does not exceed its upper bound limit, equation (3.3).

iii) any experimentally measured fluxes are constrained via upper and lower bounds, equation (3.5), where v_m is an experimentally measured flux value.

$$\begin{bmatrix} \boldsymbol{S}_i \\ \boldsymbol{S}_e \end{bmatrix} \cdot \boldsymbol{v} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{v_m} \end{bmatrix} \tag{3.2}$$

$$\boldsymbol{v} \geq 0 \tag{3.3}$$

$$\boldsymbol{v} \leq \boldsymbol{v}_{max} \tag{3.4}$$

$$v_{m,min} \leq v_m \leq v_{m,max} \tag{3.5}$$

The next step in FBA is to define a phenotype in the form of a biological objective that is relevant to the metabolic network being studied [77]. For example, if growth is being predicted the objective is biomass production. This is the rate at which metabolites are converted to biomass constituents such as proteins and lipids. Biomass production can be represented by the addition of an artificial biomass reaction that encompasses all reactions that lead to growth of the cell. The objective function is maximised, equation (3.6), where Z is the objective function and δ_i is a binary variable indicating if the reaction is on or off in the objective function.

$$max: Z = \delta_1 v_1 + \delta_2 v_2 + \cdots \delta_i v_i \quad (i = 1, .., N_R) \tag{3.6}$$

A major drawback of FBA is the assumption of optimal behaviour of the cell. Therefore, the optimal solution found may not correspond to the actual flux distribution within the cell. Therefore, when using FBA, it is vital to hypothesise that a) the cell has evolved to achieve optimal behaviour b) the objective of the cell is known and c) the objective of the cell can be transformed into mathematical form [78]. Another drawback is the restriction to one optimal flux mode e.g. biomass growth [31]. Overall, it can be said that FBA is useful at obtaining flux values within optimal conditions. However, to back up calculated values it would be necessary, and widely encouraged to perform experimental tests.

3.2.1 *Exactly Determinable Chinese Hamster Ovary Cell*

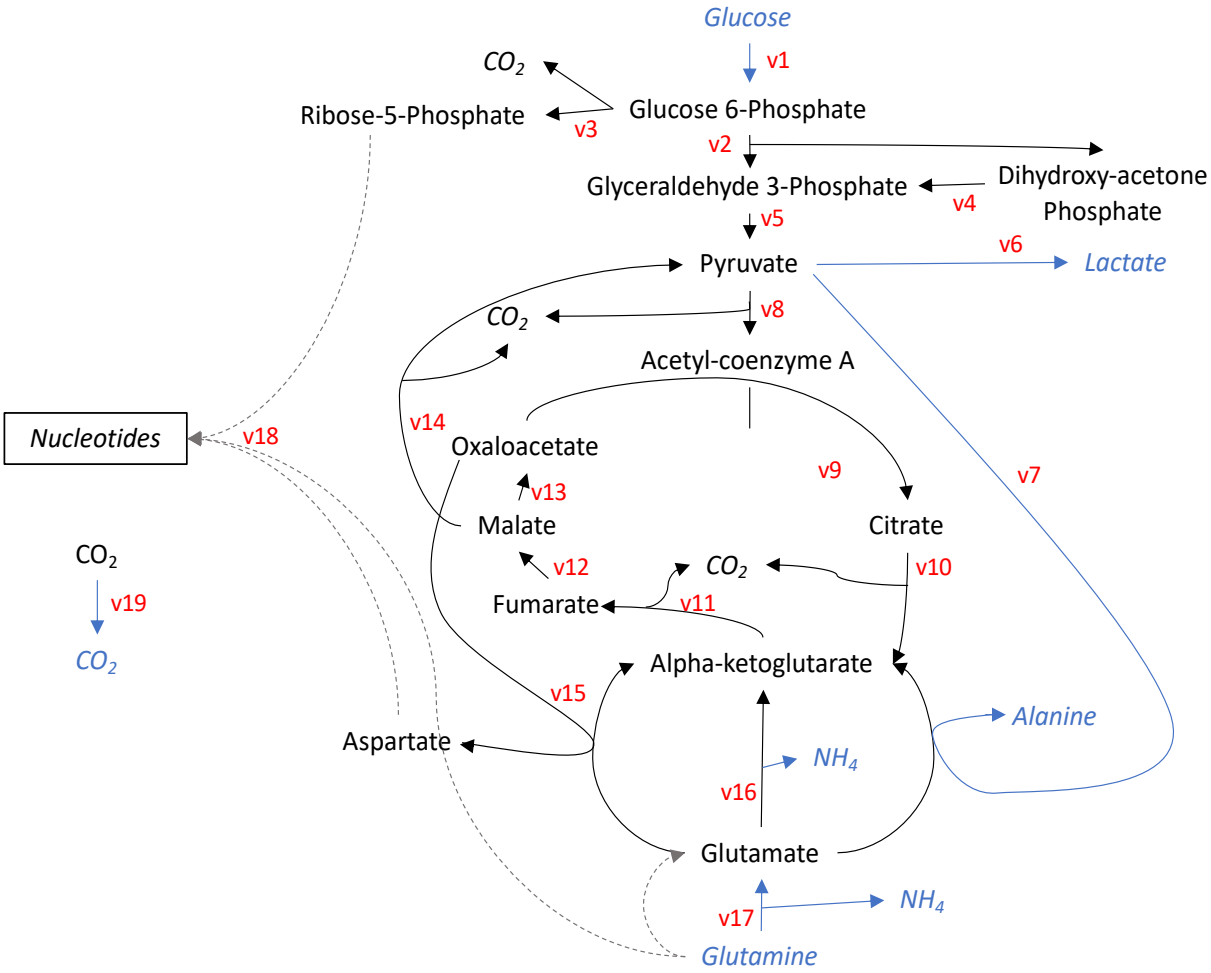

Figure 3.1 Simple CHO cell network. Blue indicates the reaction or metabolite is extracellular and black is extracellular [1]

A simple, exactly determinable Chinese hamster ovary (CHO) cell has a total of five products. These products were maximised via FBA code with flux limits of $0 \leq v \leq 10$ mM (d x 10^9 cells)$^{-1}$. These limits were chosen as the maximum flux achieved for this network was 8.1524mM (d x 10^9 cells)$^{-1}$) (Chapter 4). The products were:

1. Lactate via reaction 6
2. Alanine via reaction 7
3. Ammonia via reactions 16 and 17
4. Carbon dioxide via reaction 19
5. Biomass via reaction 18

Biomass is not explicitly defined in this network, however, purine and pyrimidine nucleotides are necessary simultaneously to produce biomass [79]. Therefore, it can be inferred that by maximising the production of these nucleotides you are increasing the biomass yield.

Lactate is the predominant waste product of this network [80]. It poses a risk to cell growth and the productivity of the cell over its lifetime, particularly at high concentration during the manufacturing of recombinant proteins [81]. Recombinant protein yield is a function of the cell density and protein production [82]. Lactate engagers the latter. To reduce lactate production process optimisation can be used however this is often restricted to the specific needs of the producer and their steps they take to optimise the process [81]. Genetic engineering is also another option.

Ammonia has a similar effect to lactate [81, 83]. The effect of ammonia on the metabolism of glucose, glutamine, and other amino acids in a batch culture of recombinant CHO cells have been investigated. The yields of cells to glucose, glutamine and other consumed amino acids decreased with the increase of initial ammonia concentrations. The metabolic pathways taken changed when ammonia concentrations were higher [84]. The glucose consumption was more prone to form lactate by anaerobic metabolism; thus, creating more 'waste' and inhibiting the cell further.

Alanine is an amino acid that is used to generate the proteins, often required in therapeutic recombinant protein production. Alanine is a non-essential amino acid; however, it has been found to have a positive impact on the biomass production in the CHO cell. Inclusion of all amino acids regardless of their assumed essentiality leads to greater biomass production [85]. Overall, it is important to maximise alanine and biomass production in this network, whilst trying to minimise lactate and ammonia.

3.2.1.1 Presenting Flux Data

There are many ways to present flux data, for example in a tableau or heat map. However, this work proposes that the use of digraphs to display flux is more meaningful and accessible. Figure 3.3 to Figure 3.5 provide the bipartite representations of the reactions and their associated flux, in mM (d x 10^9 cells)$^{-1}$, required to maximise the 5 products. The numbers

shown between reactions and metabolites are the associated fluxes. For example, in Figure 3.3, reaction 1 which consumes glucose and produces Glucose-6-P has a flux of 5 mM (d x 10^9 cells)$^{-1}$. Both ammonia and CO_2 are maximised from the same set of reactions shown in Figure 3.4. Based upon the biological significance of the products reaction 6, producing lactate, and reactions 16 and 17, producing ammonia should be avoided.

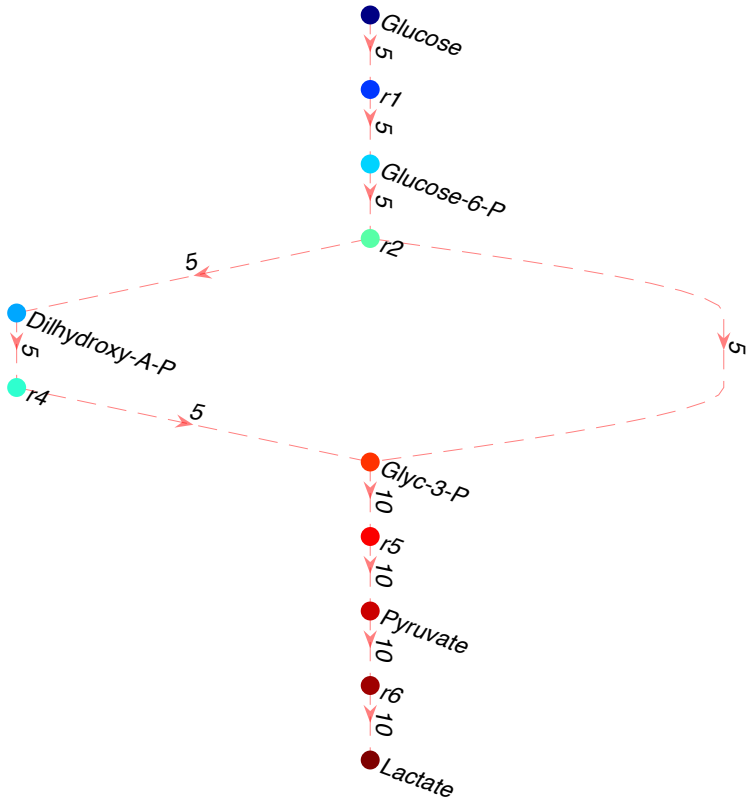

Figure 3.3 FBA results to maximise lactate production

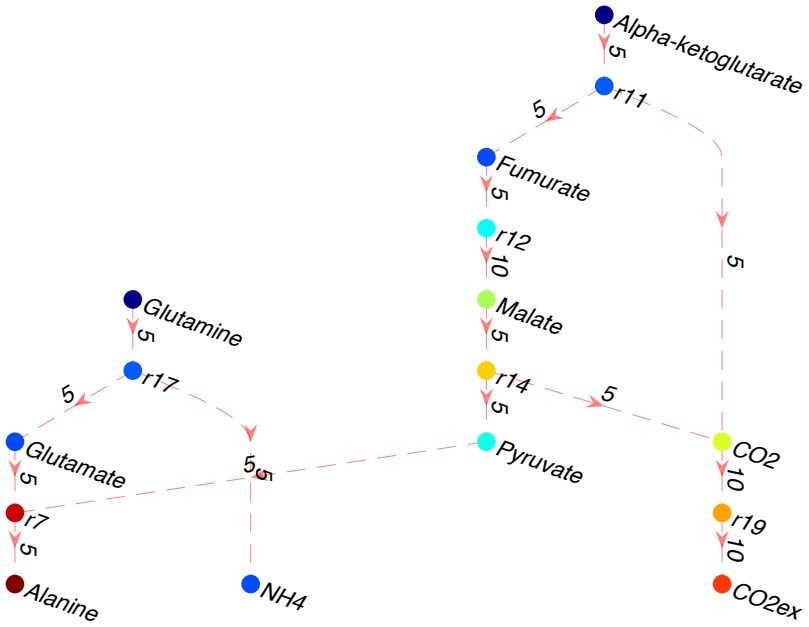

Figure 3.2 FBA results to maximise Alanine production

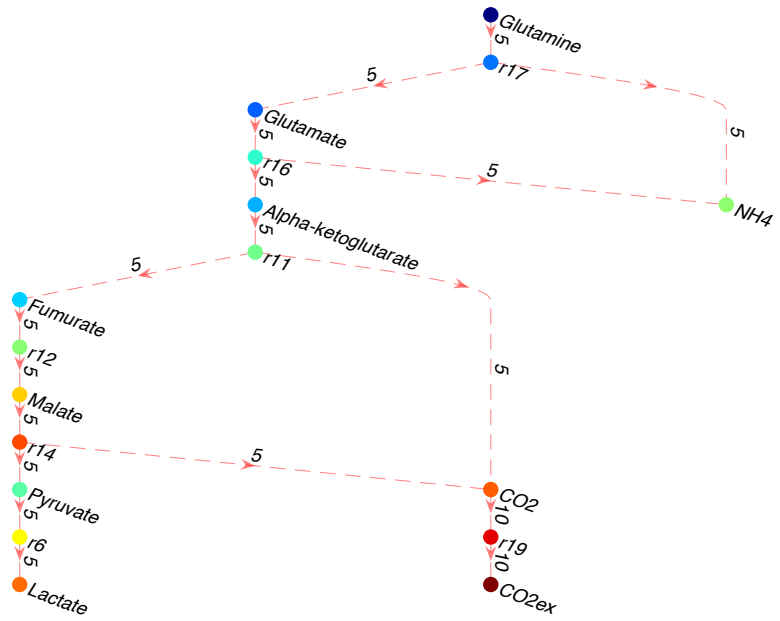

Figure 3.4 FBA results to maximise ammonia or CO^2 production

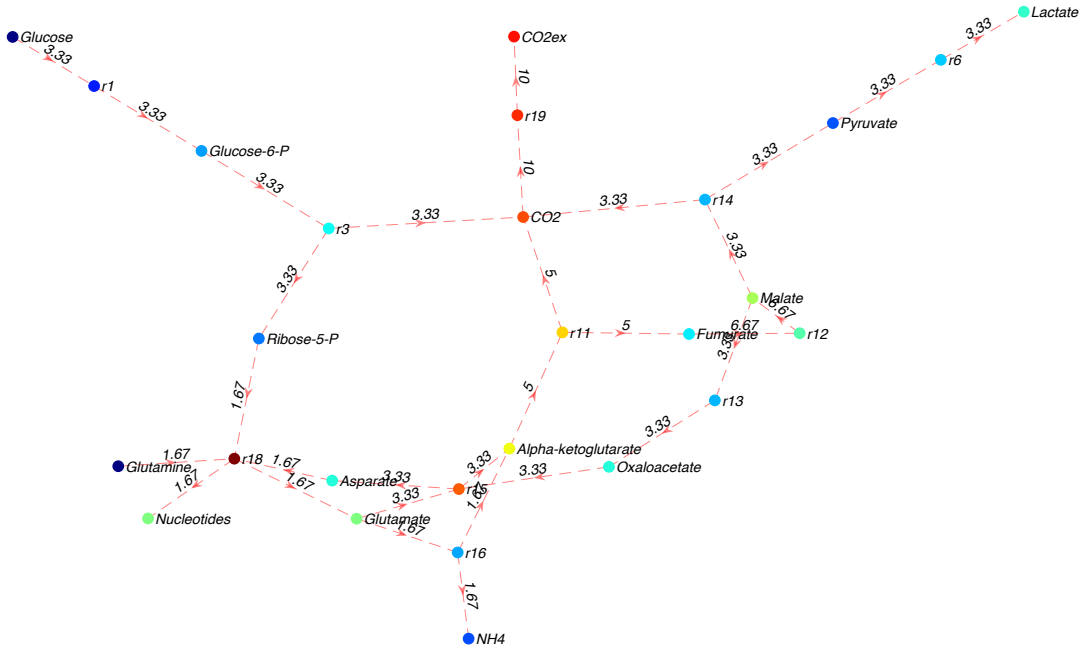

Figure 3.5 FBA results to maximise biomass production

Examination of Figure 3.2 highlights the trade off in maximising alanine is the production of ammonia via reaction 17. Figure 3.4, however, shows that the maximisation of undesirable ammonia will also produce another undesirable product, lactate. Therefore, FBA in this case shows that maximising the production of desirable products will produce some undesirables but does not maximise them. This example shows the use of FBA in learning how to drive production without comprising on cell growth.

3.3 Flux Variability Analysis

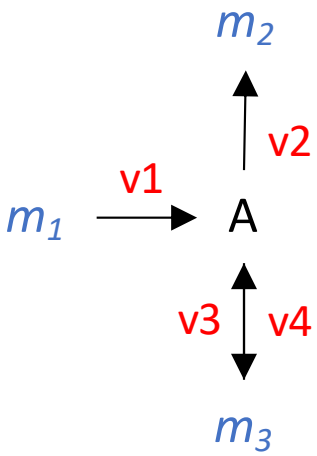

Figure 3.6 Example network to demonstrate stoichiometric representation

Every flux within a network can be defined as a convex combination of elementary flux modes (EFMs),

$$\boldsymbol{v} = \sum_{i=1}^{nEFM} \boldsymbol{\mu_i e_i} = \mathbf{E\mu} \tag{3.7}$$

This equation is equivalent to the weighting's equation presented in Chapter 2, equation (2.13).

Figure 3.6 shows a simple network with 3 EFMs,

$$\boldsymbol{E} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \tag{3.8}$$

These three EFMs define a polyhedral cone, which contains all possible flux distributions. If all reaction fluxes are unknown, $\boldsymbol{S_N}$, i.e., none have been measured, $\boldsymbol{S_m}$, then the system is underdetermined as equation (3.1) is true. In any underdetermined case then the basis search space for EFMs is,

$$\begin{bmatrix} \boldsymbol{S_N} & \mathbf{0} \\ \boldsymbol{S_m} & -\boldsymbol{v_m} \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{v} \\ \mathbf{1} \end{bmatrix} = \mathbf{0} \tag{3.9}$$

which leads to two extreme rays,

$$\boldsymbol{e_1} = [1\ 0\ 1\ 0] \tag{3.10}$$

$$\boldsymbol{e_2} = [0\ 1\ 0\ 1] \tag{3.11}$$

These rays are equivalent to two of the EFMs in equation (3.7); $\boldsymbol{E_1}$ exists within the space created by these rays. The extreme rays can be used to generate a flux spectrum for the reaction network,

$$\begin{bmatrix} 0 \leq v_1 \leq 1 \\ 0 \leq v_2 \leq 1 \\ 0 \leq v_3 \leq 1 \\ 0 \leq v_4 \leq 1 \end{bmatrix} \tag{3.12}$$

Any solution within these bounds is a feasible flux for the network. Figure 3.7 provides the axis representation of the extreme rays to define the flux distribution space [86]. An alternative to this procedure is FVA, to generate the same spectrum of results [87]. The spectrum of results obtained are the maximum and minimum feasible flux values for a specific objective function, which is often biomass [88]. There do exist variations on this method, fast

thermodynamically constrained flux variability analysis (tFVA) for example. This method removes thermodynamically infeasible reactions to reduce solver time [36].

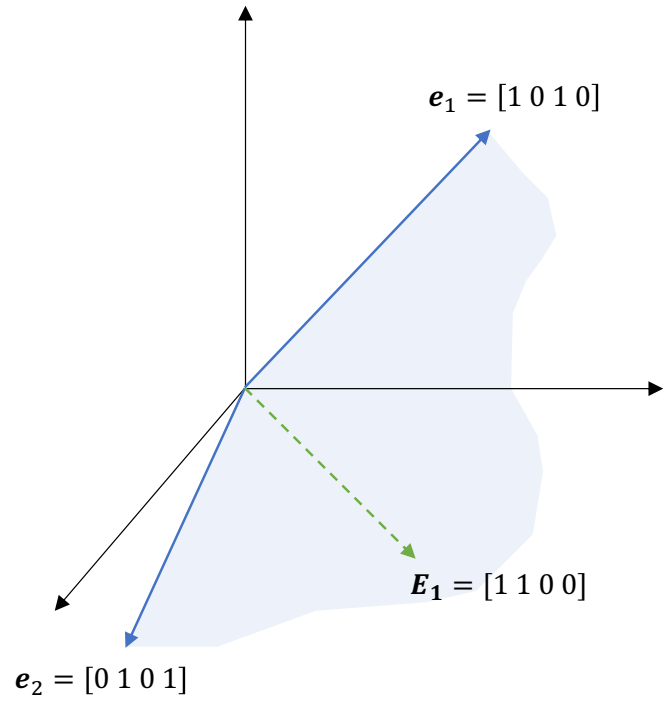

Figure 3.7 Axis depiction of extreme rays and elementary flux mode

To perform FVA, an optimal value is determined for one or a combination of reactions to maximise an objective. Then the possible maximum and minimum flux values are estimated, whilst ensuring the objective function is met. The set of flux ranges found define the boundary of the optimal solution space, for a particular objective [89]. Equations (3.2), (3.3) and (3.5) are used along with setting the objective function to be the maximum possible solution, which is the flux for a specified reaction. All fluxes must be maximised and minimised to ensure the objective function is met, equations (3.13) and (3.14).

$$max: Z = v_i \qquad (i = 1, .., N_R) \tag{3.13}$$

$$min: Z = v_i \qquad (i = 1, .., N_R) \tag{3.14}$$

3.3.1 Underdetermined Chinese Hamster Ovary Cell

FVA was used on an underdetermined CHO network from Provost *et al*'s work, Figure 3.8, using the undesirable lactate production as the objective function [90]. This network is the same as that presented in Figure 3.1 but expanded to encompass the pentose phosphate pathway. Lactate is maximised via flux through reaction 6. Due to the lack of comprehensive data on CHO cell composition, it is acceptable to not use growth of biomass as the objective [91]. The minimum flux values of 0 in Figure 3.8 are expected due to the irreversibility constraint. It is also expected that reaction 6 will have the same value for minimum and maximum flux as this reaction solely produces lactate and is being maximised. In Provost *et al*'s work, FVA was not performed as the exactly determinable network was used. Therefore, the results in Figure 3.8, are not comparable against literature. However, based upon typical flux values expected in the CHO cell, see Chapter 4, from Provost *et al*'s data, the flux ranges found are acceptable. Any flux within the flux range can occur. Figure 3.8 results emphasise that only 5 reactions need to be operational to maximise the yield of lactate. As discussed previously, lactate will hinder the growth of the CHO cell. The flux ranges achieved by FVA highlight the ease with which lactate can be largely produced. Therefore, FVA offers insight into ease of production through a network.

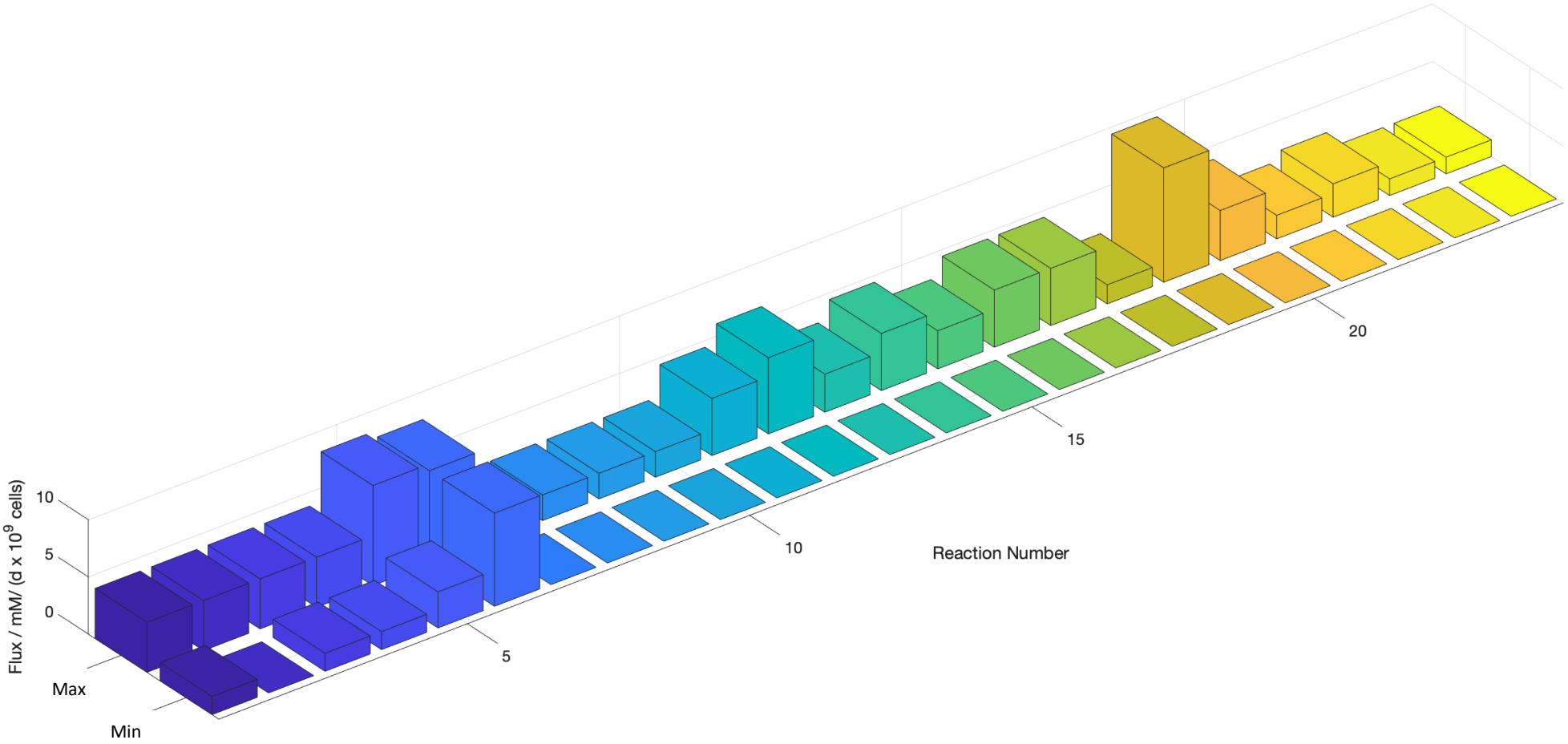

Figure 3.8 Maximum and minimum fluxes for underdetermined CHO cell network

3.4 Conclusion

For undetermined networks both FBA and FVA can be performed. FBA requires experiments to confirm the results but highlights key reactions in the production of desirable and undesirable metabolites. It can be used to learn how to drive production without compromising on cell growth. FVA produces a range of fluxes with the aim of maximising reactions. It gives an insight into how easy it is to maximise metabolites at a minimum flux with as few reactions as possible. This has been found to be particularly useful when network routes that produce undesirable products at a low overall flux distribution. Both FBA and FVA offer an insight in a cell's flux distribution but with experimental data true flux estimates can be achieved. The next chapter will discuss how this experimental data can be used. Integration of these analysis techniques into *E. coli* EFM detection will be investigated in Chapter 7.

Chapter 4 Integrated Metabolic Flux Analysis

4.1 Introduction

Metabolic flux is the passage of a metabolite through a reaction system over time. Metabolic flux analysis (MFA) uses experimental measurements for extracellular metabolites to estimate unknown intracellular metabolite movement through the reaction system. MFA is only possible for fully determinable systems [78]. A system is fully determined if,

$$\mathrm{rank}(\boldsymbol{S_N}) = \mathrm{u} \tag{4.1}$$

MFA, based on stoichiometric equations representing reactions within metabolic networks, is widely used to determine the metabolic flux distribution that reflects or represents cell physiology [92]. It creates a metabolic map revealing the contribution of each reaction to the overall metabolic processes of substrate consumption and product formation [78]. However, using MFA has several drawbacks: 1) it cannot be used in undetermined systems, as defined by Klamt *et al* [76]. This case is often encountered due to the lack of measurable fluxes. 2) Errors in measurements introduce noise into the analysis, leading to unreliable flux estimations [93]. 3) Reversible reactions cannot be examined as the flux estimated would be the net value for that particular reaction [94].

The work presented in this chapter aims at improving the effect of sampling and errors within this analysis. It was proposed to use an integrated form of metabolic flux analysis (iMFA) as initially proposed by Portela *et al* [93]. However, the work presented here goes further by discussing how observations of material change via the intracellular routes can be made. As will be demonstrated, observation of material change allows for constants in reaction rates to be estimated and a dynamic simulation produced [95, 96]. The method was applied to two simulations; with continuous, batch and fed-batch conditions examined [90, 93].

4.2 Metabolic Flux Analysis

4.2.1 Methods

In a bio-reactor the rate of change of mass (or concentration) of the external species (substrates or products) can be defined by [97],

$$\frac{d\boldsymbol{m}}{dt} = \frac{d(\boldsymbol{c}V)}{dt} = \boldsymbol{S}_e \cdot \boldsymbol{v}(\boldsymbol{t}) + F \cdot \boldsymbol{c_F} \tag{4.2}$$

In this equation, $\boldsymbol{m} = (m_1, \quad \cdots \quad m_m)^T$ is a vector of the mass (g) of the individual metabolites, $\boldsymbol{c} = (c_1, \quad \cdots \quad c_m)^T$ a vector of the concentration $(g.l^{-1})$ of the individual species, V is volume (l), $\boldsymbol{c_F} = (c_{F,1}, \quad \cdots \quad c_{F,m})^T$ a vector of the concentration of the external species in a feed stream $(g.l^{-1})$ and F $(l.h^{-1})$ the feed rate (a batch reactor has $F = 0$).

The basis of MFA is the assumption that there is no accumulation of intracellular metabolites. As discussed in Chapter 2, Elementary flux modes (EFMs) represent non-decomposable metabolic pathways between the substrates and final products. Knowledge of the EFMs allows the time dependent vector of fluxes to be written as a non-negative combination of specific reaction rates, $\boldsymbol{r}$ $(m_{ex} \times 1)$,

$$\boldsymbol{v}(t) = \boldsymbol{E} \cdot \boldsymbol{r}(t) \cdot V \tag{4.3}$$

In this equation, $\boldsymbol{E}$ $(\boldsymbol{e} \times N_R)$ is a matrix of EFMs and $\boldsymbol{r}(\boldsymbol{t})$ is a vector of the specific reaction rates $(g.l^{-1}h^{-1})$. $\boldsymbol{r}(t)$ can be modelled, for example, via the generalized Michaelis-Menten rate expression [90, 96],

$$r(t) = \frac{X \cdot a \cdot c_i}{K + c_i} \tag{4.4}$$

In this equation, a is the maximum rate achieved by the system, X is biomass concentration and K is the Michaelis constant.

Substituting equation (4.2) into equation (4.3) gives,

$$\frac{d\boldsymbol{m}}{dt} = \frac{d(\boldsymbol{c}V)}{dt} = \boldsymbol{S}_e \cdot \boldsymbol{E} \cdot \boldsymbol{r}(t) \cdot V + F \cdot \boldsymbol{c_F} = \boldsymbol{S}_K \cdot \boldsymbol{r}(t) \cdot V + F \cdot \boldsymbol{c_F} \tag{4.5}$$

The matrix $\boldsymbol{S}_K = \boldsymbol{S}_e \cdot \boldsymbol{E}$ is the macroscopic stochiometric matrix, relating the external substrates and products. Expanding the differential this may also be written as,

$$\frac{d\boldsymbol{c}}{dt} = \boldsymbol{S}_e \cdot \boldsymbol{E} \cdot \boldsymbol{r}(t) + \frac{F}{V} \cdot (\boldsymbol{c_F} - \boldsymbol{c}) \tag{4.6}$$

Using equation (4.7) we can define equation (4.8),

$$\begin{bmatrix} \boldsymbol{S}_i \\ \boldsymbol{S_e} \end{bmatrix} \cdot \boldsymbol{v}(t) = \begin{bmatrix} 0 \\ \boldsymbol{v_m}(t) \end{bmatrix} \tag{4.7}$$

$$\boldsymbol{v_m}(t) = \boldsymbol{S_e} \cdot \boldsymbol{v}(t) = \frac{d\boldsymbol{m}}{dt} - F \cdot \boldsymbol{c_F} = \frac{d(\boldsymbol{c}V)}{dt} - F \cdot \boldsymbol{c_F} \tag{4.8}$$

For an exactly determinable system equation (4.8) can be solved subject to the constraint, $\boldsymbol{v} \geq \boldsymbol{0}$. Note, however, given experimental measurements the method requires an approximation of the derivative terms in equation (4.8). Therefore, by approximating the derivative using measured concentrations you can also get these specific rates over time. Least squares, linear or quadratic programming can be used to get the flux value at each time point provided the system is not underdetermined.

When using experimental data, it is necessary to know, or estimate, the extracellular fluxes – fluxes that cross the cell boundary. This can be done via differentiation or integration, however, the use of differentiation is most commonly used [98]. To use the differential method, a function is fitted to approximate the measured extracellular metabolite (metabolites classed as outside the cells wall, Chapter 2) concentrations, then the derivative of this function is found [99]. The derivative is then divided by the biomass concentration at that specific time, to estimate extracellular fluxes. It is known that this method can amplify errors derived from the measured concentrations due to poorly estimated derivatives.

4.2.2 Intracellular Flux Estimation

Constrained optimisation is used to estimate fluxes in MFA due to measurement errors present within experimental data. Dai and Locasale discussed using large-scale constrained non-linear least squares set up to evaluate unknown fluxes. This is done via the minimisation of the difference in simulated isotopomer distribution profiles (technique to measure synthesis of biological polymers *in vivo* [100]) from assumed fluxes and the experimentally measured isotopomer distribution profiles [101].

As opposed to using a least squares objective function (minimising the squared error between an output and a predicted output) in this work the objective function used was the sum of the absolute errors (the ℓ_1 norm), known as least absolute deviations (LAD). Unlike least squares, LAD is not as sensitive to outliers and produces robust flux estimations [102]. This is due to not squaring the residual, like with least squares. Therefore, the following cost function was minimised,

$$J = \|\boldsymbol{S}_e\, \boldsymbol{v}(t) - \boldsymbol{v_m}(t)\,\|_1 \tag{4.9}$$

The advantage of using the ℓ_1 norm is that it may be formulated as a linear objective function allowing the use of linear programming methods. Rewriting equation (4.9) using a vector of artificial variables, $\boldsymbol{z} = (z_1, \dots, z_n)^{\mathrm{T}}$ where n is the number of reactions, gives,

$$J = \sum_{i=1}^{n} z_i \tag{4.10}$$

Equation (4.10) is minimised subject to the constraints,

$$-\boldsymbol{z} + \boldsymbol{S}_e\boldsymbol{v}(t) \leq \boldsymbol{v_m} \tag{4.11}$$

$$-\boldsymbol{z} - \boldsymbol{S}_e\boldsymbol{v}(t) \leq -\boldsymbol{v_m} \tag{4.12}$$

$$\boldsymbol{z} \geq 0 \tag{4.13}$$

$$\boldsymbol{S}_i \boldsymbol{v}(t) = \boldsymbol{0} \tag{4.14}$$

Equations (4.11) and (4.12) minimise the sum of the errors generated by the extracellular measured fluxes. Equation (4.13) ensures that the error term is positive, and equation (4.14) is the steady state constraint.

4.2.3 Hypothetical Cell Example

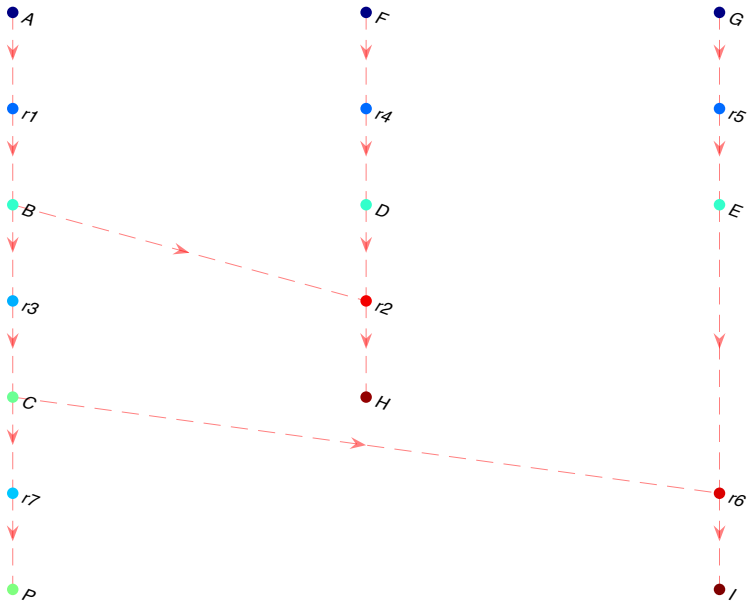

Figure 4.1 Hypothetical simple cell metabolic network

Figure 2.10 shows the digraph of a simple cell network consisting of three extracellular substrates, A, F and G and three extracellular products H, I and P [93]. The stoichiometry for this network is,

$$\boldsymbol{S_i} = \begin{bmatrix} 1 & -1 & -2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 & -1 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \tag{4.15}$$

$$\boldsymbol{S_e} = \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{4.16}$$

As this network is hypothetical reaction rates had to also be generated to simulate the network. There exists three EFMs for this network, equation (4.17), which can be used alongside extracellular stoichiometry to find the macroscopic network $\boldsymbol{S_K}$, equation (4.18).

$$\boldsymbol{E} = \begin{bmatrix} 1 & 2 & 2 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \tag{4.17}$$

$$\boldsymbol{S_K} = \boldsymbol{S_e} \cdot \boldsymbol{E} = \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 & 2 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} -1 & -2 - & 2 \\ 0 & 0 & -1 \\ -1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \tag{4.18}$$

Three reaction rates for the macroscopic stoichiometry, $\boldsymbol{S_K}$, were used to generate the rates and therefore the measured fluxes for this network. As no biomass was defined for this network it is assumed that metabolite 'I' is biomass. The three rates were,

$$r_1 = \frac{0.4 \cdot c_A \cdot c_I}{c_A + 0.1 + 10 \cdot c_H^2} \tag{4.19}$$

$$r_2 = \frac{0.2 \cdot c_A \cdot c_I}{c_A + 4.5} \tag{4.20}$$

$$r_3 = \frac{0.1 \cdot c_G \cdot c_I}{c_G + 10.5} \cdot \frac{c_A}{c_A + 0.001} \tag{4.21}$$

Using equation (4.6) this allowed the set of ordinary differential equations (ODEs) to be integrated to obtain the extracellular species concentrations with respect to time. The equations were integrated with *ode45* – sampled every 10 seconds. Integration was performed with ode45 due to the non-stiff nature of the equations.

If measured fluxes were unknown then derivative approximation was required; therefore, polynomial fitting can be performed. The ideal data (no noise), which is often not available, is used to ascertain the degrees of the polynomial equations expected for each extracellular metabolite as a starting point. For MFA the 'polyfit' and 'polyder' functions in MATLAB2021a were used to estimate the derivative at measured concentrations. Equations (4.22) and (4.23) give an example of what these equations may look like, where b,d,e and f are constants.

$$c = bt^3 + dt^2 + et + f \tag{4.22}$$

$$\frac{dc}{dt} = 3bt^2 + 2dt + e \tag{4.23}$$

4.2.4 Exact Derivatives

For an ideal system only the substrate stochiometric matrix is required to ensure the system is exactly determinable [76]. The LAD formulation of MFA gives the results shown Figure 4.2. For example, the flux in reaction 1 is initially at $0.1 g.l^{-1}.h^{-1}$ rising to $0.9 g.l^{-1}.h^{-1}$ over 100 hours of cell life. Each time point has an associated flux for each reaction of the network. As the system is modelled as batch, without substrate formation as a product, the concentrations of substrates will decrease over time resulting in a decrease in flux to a steady state.

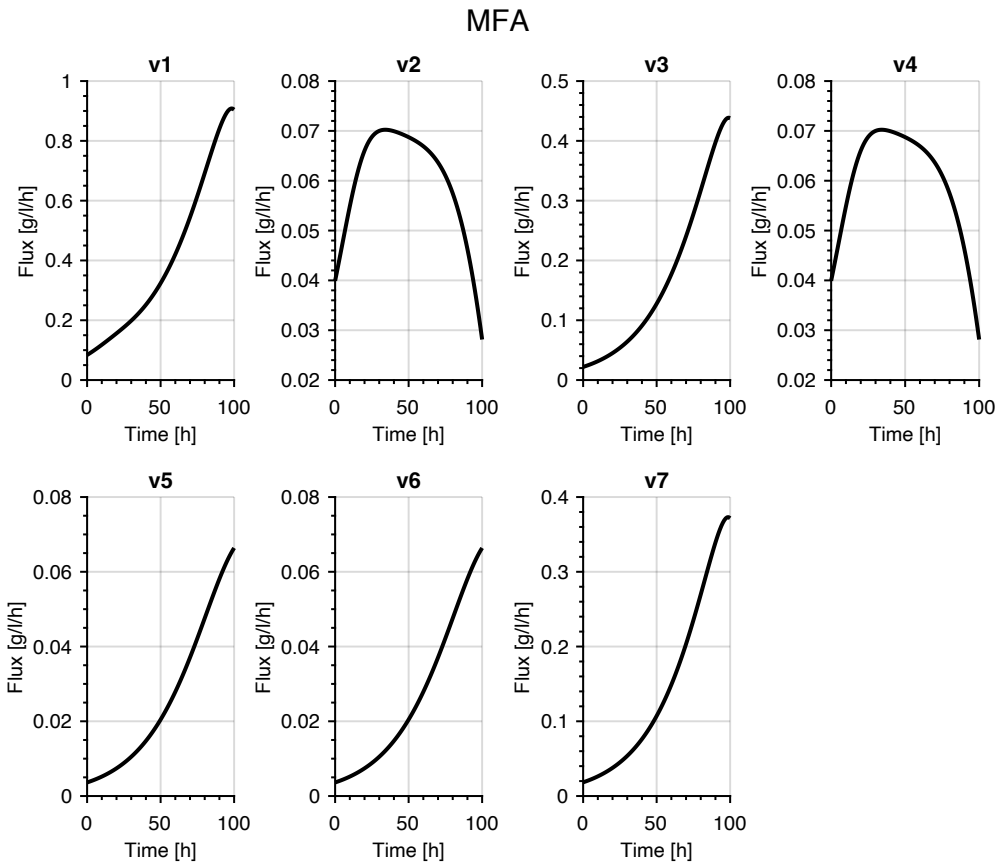

Figure 4.2 MFA results for all fluxes in a hypothetical cell from 0 to 100 hours

4.2.5 Noisy Derivatives

To simulate realistic conditions noise is required. Random noise was added to each concentration data point for all extracellular metabolites. Normal distributed noise is used with a standard deviation of [0.8713, 0.4227, 1.2240, 0.9400, 1.4182] for each species respectively. A total of 2.5% of the random noise is added onto all of A's data points, 0.25% of the random noise to G, F, H, and I and 1% of the random noise to P. The randomised noise

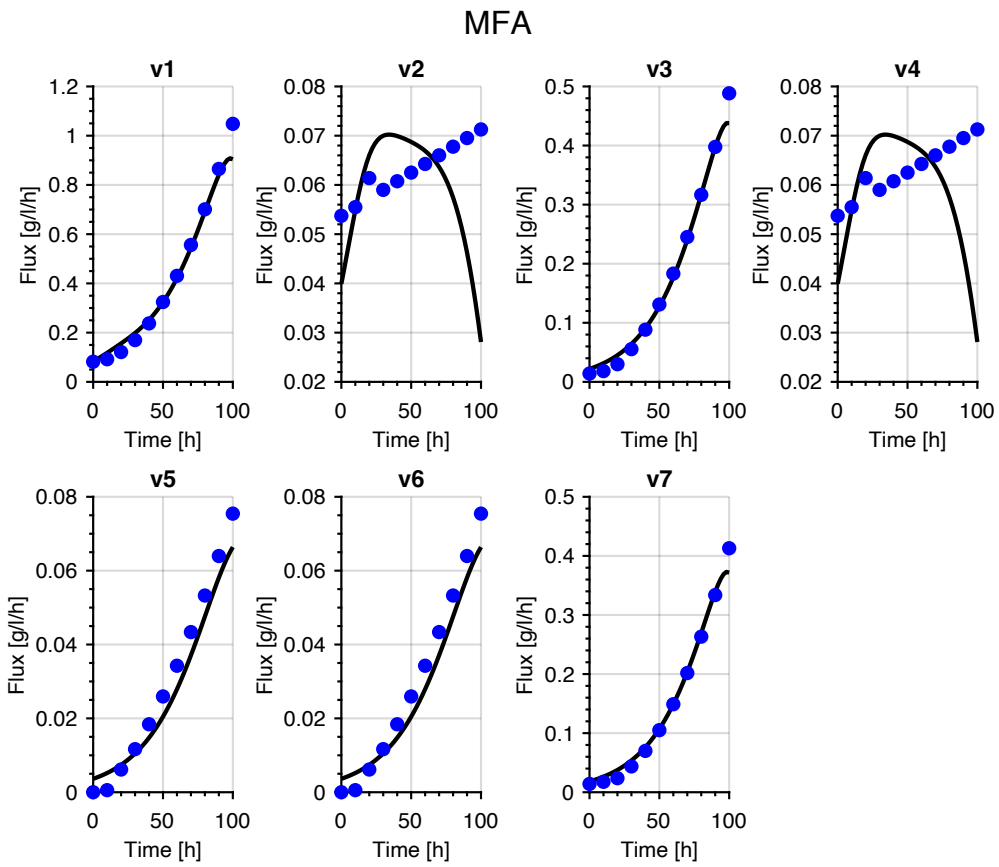

Figure 4.3 MFA ideal flux = black line, polynomial fitting = blue dot

is kept constant. There is no noise at the initial data point.

MFA was trialled with derivative estimated measured fluxes using the noisy data. Figure 4.3 shows the ideal data (no noise) compared to the flux found via polynomial fitting (noisy data used). The polynomial fitting data has been sampled every 10 hours. The fit is poor if the flux value plateaus and decreases over the simulation time. This is often the case if multiple phases of cell life are modelled. Flux through reactions often decrease in the death phase. The polynomial fitting accounts only for growth, and therefore, a gradual increase of flux over time. The polynomial equation required for reactions 2 and 4 are different after 35 hours. To

improve the fit better estimations of the equations would be required; a process that could be lengthy as the number of reactions in a network increase. It would not be possible to make an accurate estimation of flux using this method if all reactions required highly accurate polynomial's, particularly with high likelihood of noise in any real data set [95].

4.2 Integrated Metabolic Flux Analysis

4.3.1 Methods

iMFA uses an integrated form of equations (4.5) and (4.7) to remove the need for estimating the derivative of concentration (or mass) with respect to time. Therefore, specific measured rates are not needed; these rates can be hard to determine so are often not available [93]. The integrated form of equation (4.5) is,

$$\begin{aligned} \boldsymbol{m}(t) &= \boldsymbol{m}(0) + \int_0^t (\boldsymbol{S_e} \cdot \boldsymbol{v}(t) + F \cdot \boldsymbol{c_F})\, dt \\ &= \boldsymbol{m}(0) + \boldsymbol{S_e} \int_0^t \boldsymbol{v}(t) dt + \int_0^t F \cdot \boldsymbol{c_F} dt \end{aligned} \tag{4.24}$$

The steady state assumption is,

$$\boldsymbol{S}_i \cdot \int_0^t \boldsymbol{v}(t) dt = \boldsymbol{0} \tag{4.25}$$

This is generalised in equation (4.26).

$$\begin{bmatrix} \boldsymbol{S_i} \\ \boldsymbol{S_e} \end{bmatrix} \cdot \int_0^t \boldsymbol{v}(t) dt = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{m_m}(t) \end{bmatrix} \tag{4.26}$$

In this equation $\boldsymbol{m}_m(t)$ is a vector of masses, which are obtained via re-arrangement of equation (4.8),

$$\boldsymbol{m_m}(t) = \boldsymbol{S}_e \int_0^t \boldsymbol{v}(t) dt = (\boldsymbol{m}(t) - \boldsymbol{m}(0)) - \int_0^t F \cdot \boldsymbol{c_F} dt \tag{4.27}$$

Note that for a batch reactor, equation (4.27) is defined without the need for an approximation. For a semi-batch reactor, there is the need to approximate the integral term. For an exactly determinable system equation (4.24) can be solved subject to the constraint, $\int_0^t \boldsymbol{v}(t) dt \geq \boldsymbol{0}$. This provides the integrated fluxes with units of mass.

There is no need to approximate the derivative for iMFA as integrating (4.16) with respect to time yields a predicted polynomial, example given in equation (4.28). Therefore, errors from the polynomial equation are not magnified.

$$\int_{t0}^{t1} \frac{dc}{dt} dt = bt^3 + dt^2 + et + f \tag{4.28}$$

4.3.2 Hypothetical Cell Example

4.3.2.1 Exact Derivatives

The same network used in section 4.2.2 is used again but for iMFA. The results for the seven reactions with ideal data are given in Figure 4.4.

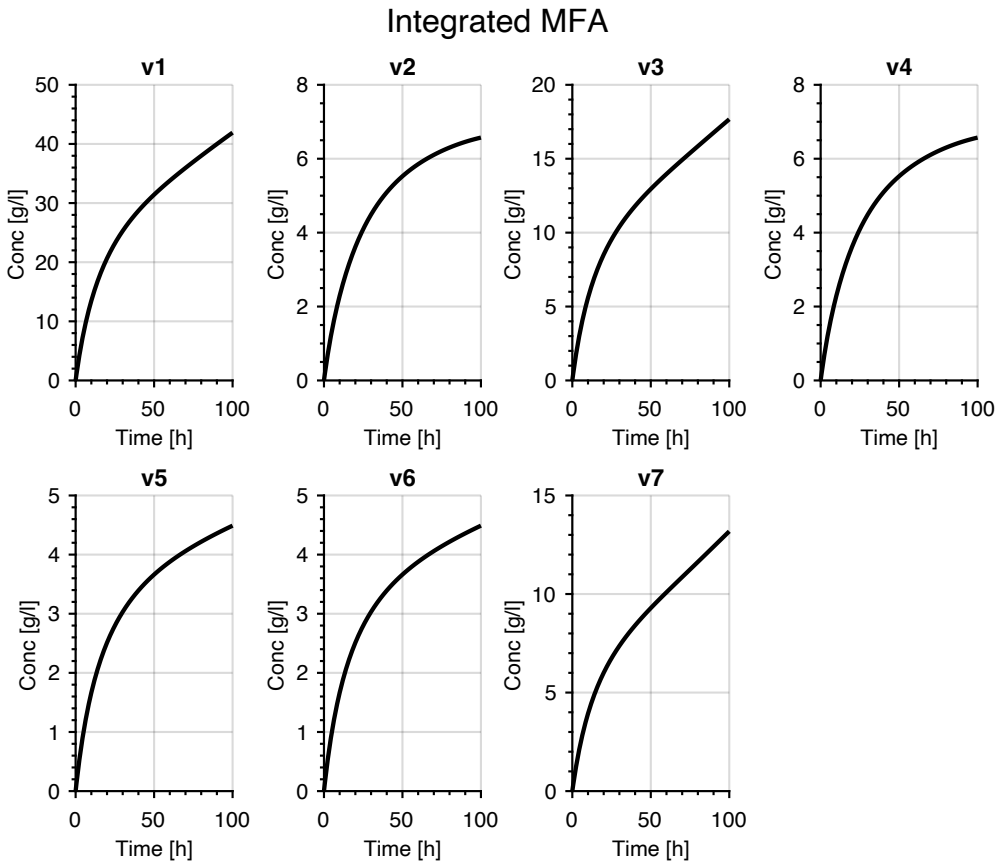

Figure 4.4 iMFA results for all fluxes in a hypothetical cell from 0 to 100 hours

The integrated form of MFA for ideal data allows the conservation relationships within the network to be examined. Each time point has an associated integrated flux, otherwise known

as concentration, Figure 4.4. The manner of which Figure 4.4 should be interpreted is through example. The plot for v_1 shows that at 100 hours of run time, a total of $40gL^{-1}$ of material has passed through the route. Only A is consumed via this route so it can be said that $40gL^{-1}$ of A has passed through the route. Moreover, as the volume is set to 1L it can be said that 40g of A has been consumed in 100 hours, as it is a substrate. This analysis can go further with the use of macroscopic material balances.

4.3.2.2 Noisy Derivatives

The same noisy data utilized in Figure 4.3 is applied to iMFA. iMFA's results are given in Figure 4.5. It is apparent that iMFA is better suited to noisy data sets than MFA; due to not requiring the derivative approximation to find the measured flux. Without estimating the derivative,

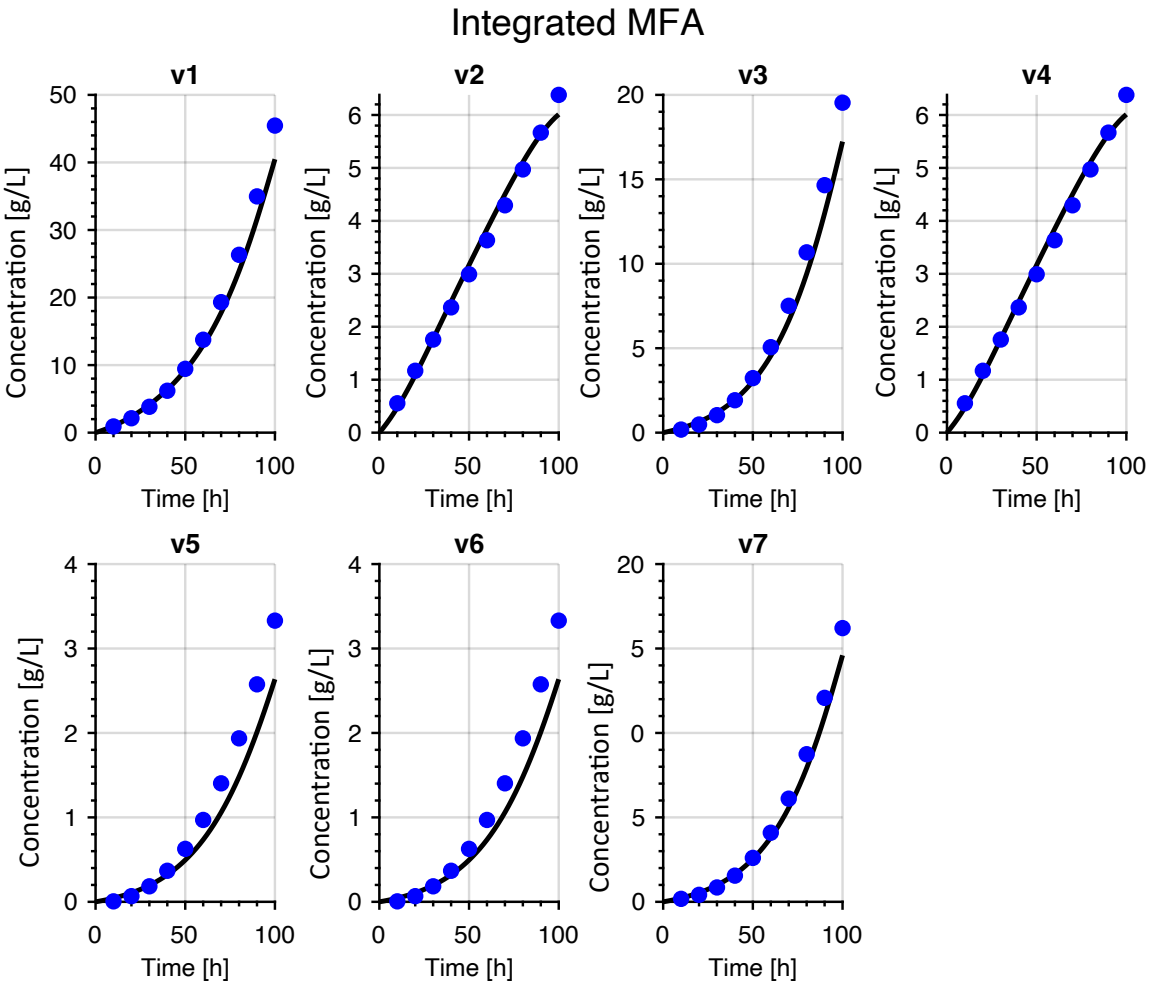

Figure 4.5 iMFA ideal flux = black line, polynomial fitting = blue dot

the error carried through the problem is reduced. Further sampling times and noise levels are tested below to give a better review on the method at determining flux.

A total of 7 sampling simulations were tested. These simulations had noise present to deviate from an ideal data set; however, the noise was maintained throughout all the tests to allow for fair comparison. The sampling times tested were 2hrs, 4hrs, 10hrs, 20hrs, 30hrs, 40hrs and 50hrs. The fit for the sample time of 10 hours from the original simulation has also been collected.

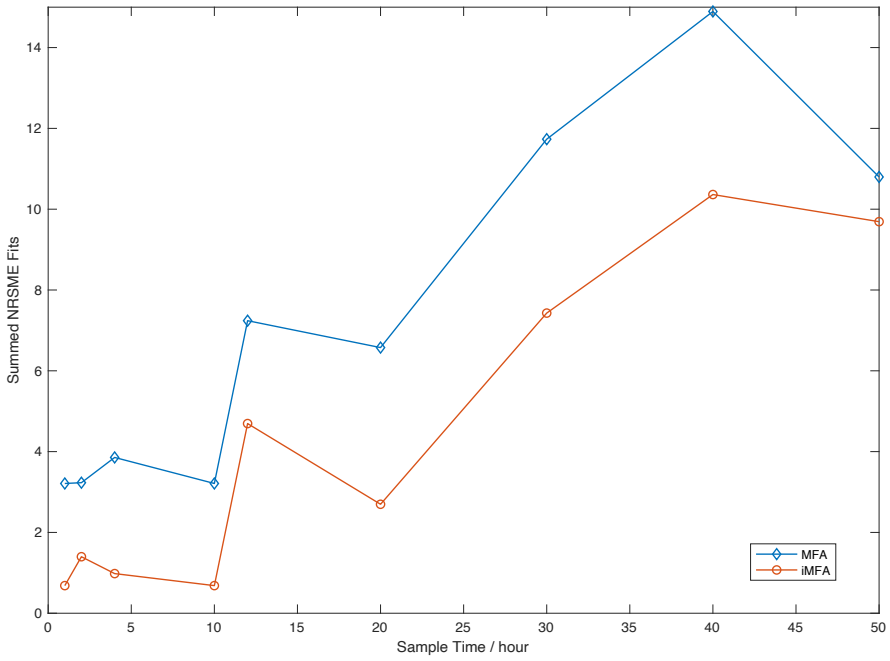

Figure 4.6 Summed NRSME fits for MFA and iMFA across the sampling times simulated

Figure 4.6 details the results obtained via sampling simulations. Normalised root-mean square error (NRMSE) was used to allow for comparison of fits across all reactions. NRMSE relates the RMSE to the observed range of the variable. Thus, the NRMSE can be interpreted as a fraction of the overall range that is typically resolved by the model. A fit closer to 0 is the ideal result. The best fit achieved was 0.68 at a 1-hour sampling time with the iMFA method. As sampling reduces the interpolated data follows the ideal data loosely, therefore, increasing error. However, the fits achieved for all iMFA simulations is an improvement on the MFA results. Overall, this method produced good fit values across the simulations tested. With

industry often sampling frequently in the day and not at night iMFA currently offers the most reliable way of estimating concentration, and therefore, material change across the cell [93].

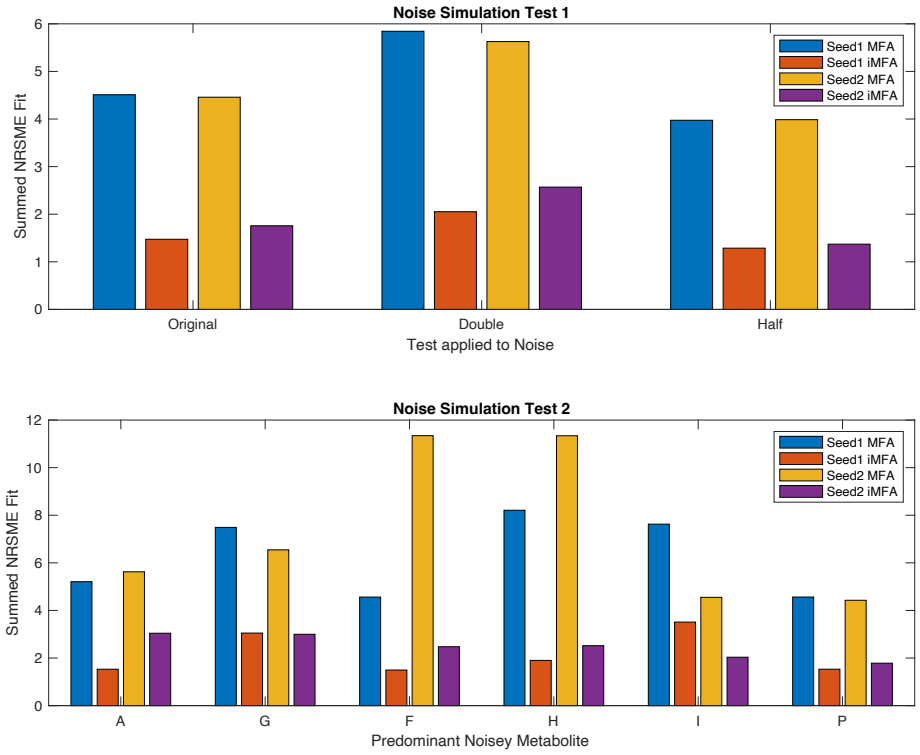

Figure 4.7 a) summed NRSME fit values for noise simulations where noise is all doubled, halved b) each metabolite largely increased individually. Metabolites from Figure 4.1.

Figure 4.7 details the fit values obtained throughout the 9 noise simulations trailed with sampling every 10 hours. Two different noise sets were applied to the data, named seed 1 and seed 2. MATLAB allows for the use of randomised numbers to be generated and reused throughout the script; this allowed for consistent random noise to be utilised. Doubling or halving the noise added did not double or halve the fit. Halving the noise improved the fit compared to the original noisy data by 26.94% for MFA and 18.2% for iMFA. All the MFA simulations yielded poorer fits when compared to iMFA. It is apparent that noise has little effect to reducing the effectiveness of iMFA. The reasoning behind this result is that the use of the derivative in MFA introduces additional error to the solution. Therefore, when noise is added this error is amplified further. iMFA does not require a derivative approximation and so any noise added to the data will not be multiplied via the derivative approximation. Figure 4.7 also presents the fit's achieved when a metabolite is set to have an error much greater than

the other extracellular metabolites. Across all metabolites, iMFA provided a better fit. Like with the doubling and halving tests, this is due to not requiring the derivative approximation in yielding a result.

4.3.3 Chinese Hamster Ovary Cell Example

The Chinese hamster ovary (CHO) cell is widely used in the pharmaceutical industry in the production of therapeutic proteins; often used in the treatment of cancer, HIV and other diseases [103]. It is necessary to generate stable CHO cell lines with optimal output to get the most out of their production [104]. Figure 4.8 gives the metabolic network for a simplified version of the CHO cell. This network has two main energetic nutrients, glucose, and glutamine. It produces lactate, Alanine, ammonia, and carbon dioxide via the four fundamental pathways: the glycolysis pathway, the glutaminolysis pathway, the TCA cycle, and the nucleotides synthesis pathway [90]. Glutamine also acts as an extracellular substrate for this network. There are 19 reactions and 21 metabolites.

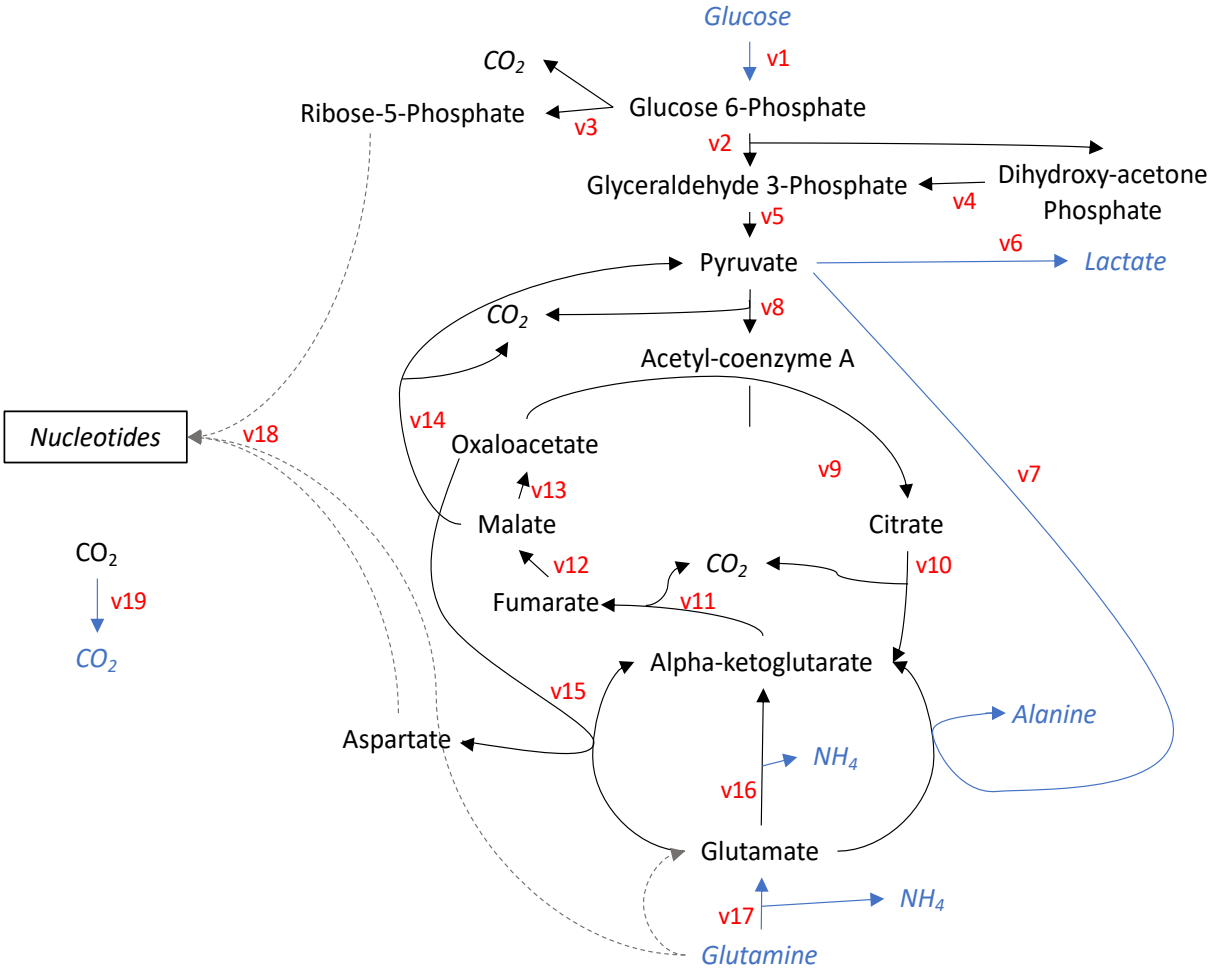

Figure 4.8 Uncompressed simple CHO cell network. Blue indicates the reaction or metabolite is extracellular and black is intracellular

The data used in this example is from Bastin *et al's* work using three CHO-320 batch cultures [1, 90]. The cells were grown with measurements taken for glucose, glutamine, ammonia, lactate, Alanine, and cell density over 80 hours, which is equivalent to the growth phase. There was available data up to 200 hours, but this encapsulated the transition and decay phases of cell life which required differing reaction rates per phase, so it was decided to use the growth phase only.

4.3.3.1 MFA and iMFA Results

MFA flux results were used to estimate concentration change over time in a simulation and this was then compared to the experimental results. The concentration results estimated from iMFA were used to generate the change in concentration over time and this was then applied to a simulation. The simulation was then compared to the experimental results. Figure 4.9

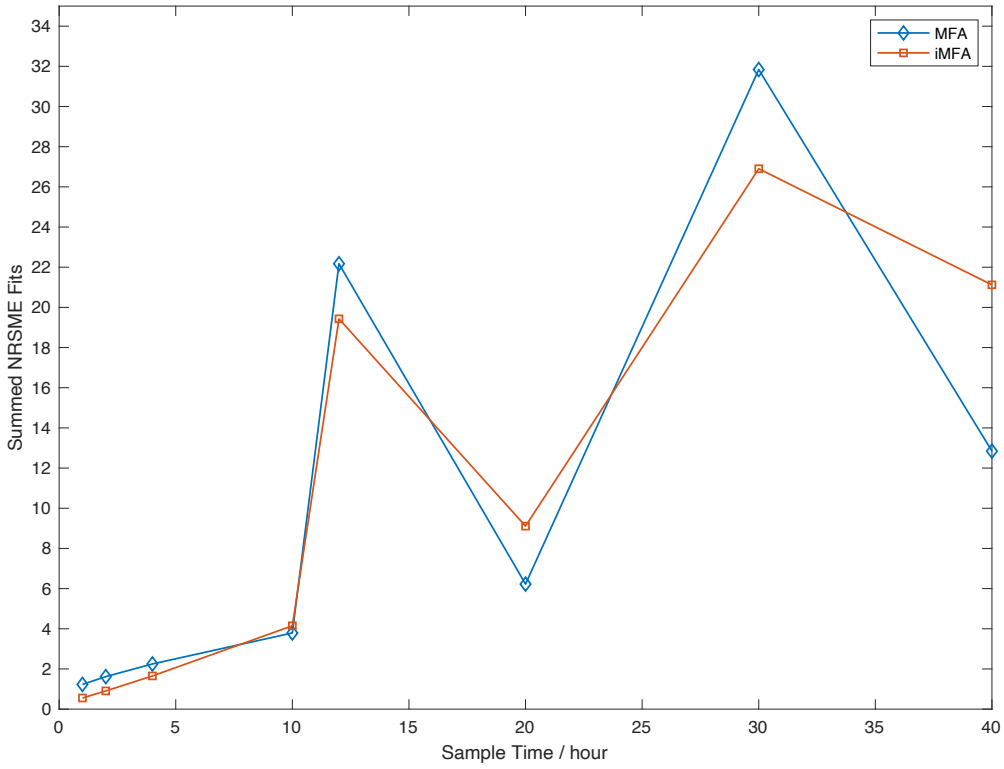

Figure 4.9 Summed fits (NRSME) achieved for the CHO cell across 7 sampling times for MFA and iMFA

shows the summed fits generated by using MFA fluxes to estimate concentration and iMFA results, achieved across all 8 sampling periods simulated. At sampling times of 10-, 20- and 40-hours MFA outperformed iMFA. However, from Figure 4.9 it can be seen this is marginal. Therefore, it can still be said that iMFA offers a reliable method compared to MFA across a wide range of sampling times. The peak and trough results are caused due to outliers present in sampling leading to poor estimates at time points.

4.3.3.2 Prediction of Rate Constants Using iMFA

As briefly mentioned above the concentration results produced by iMFA can be used to generate rates. Therefore, iMFA also offers the ability to estimate rate constants given a postulated rate model; so extracellular concentration can be predicted over the entire simulation time. Using Michaelis-Menten kinetics, reaction rates for each metabolite can be formulated. Substrate's glucose and glutamine in the CHO network are generated via equations (4.29) and (4.30) respectively. Equation (4.31) provides the reaction rate expression for products lactate, alanine, and ammonia. The stochiometric matrix defines the routes

whose material balance must be used to estimate the reaction rate for the metabolite. For example, for glutamine the iMFA results for route 17 and 18 must be combined as it is consumed via both routes. Non-linear least squares regression is used to find the constants for equations (4.29), (4.30) and (4.31), Table 4.1. The macroscopic matrix network for the underdetermined system is given by equation (4.32) showing how the EFMs relate the 2 substrates with the 3 products. Details on finding the macroscopic stoichiometry can be found in Chapter 2, section 2.9.

$$r_G = \frac{a_{G,rG} c_G}{K_{G,rG} + c_G} \tag{4.29}$$

$$r_Q = \frac{a_{Q,rQ} c_Q}{K_{Q,rQ} + c_Q} \tag{4.30}$$

$$r_P = \frac{a_{P,rP} c_G c_Q}{(K_{G,rP} + c_G)(K_{Q,rP} + c_Q)} \tag{4.31}$$

$$\boldsymbol{S_K} = \begin{bmatrix} -1 & 0 & 0 & 0 & -1 & -2 & -2 & -2 & -2 \\ 0 & -1 & -1 & -1 & 0 & -3 & -3 & -3 & -3 \\ 2 & 0 & 1 & 0 & 0 & 1 & 2 & 0 & 0 \\ 0 & 1 & 2 & 2 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix} \tag{4.32}$$

Table 4.1 Reaction rate constants determined via non-linear regression

Metabolite	A	K_G	K_Q
Glucose	0.1625	-0.7218	-
Glutamine	0.0354	-	-0.2078
Lactate	0.2512	0.9936	0.2788
Ammonia	0.0375	0.7816	0.1798
Alanine	0.0170	-0.1370	-0.0940

Table 4.1 shows that some of the half-saturation constants were in fact negative. This is not realistic as the half saturation is the concentration supporting an uptake rate one-half the maximum rate [105]. The data set used to determine these constants does contain errors as experimental data is used. These errors will be carried through to the nonlinear regression fitting of a and K. To alleviate these issues often the half-saturation constant is set to a value and the constant '*a*' estimated [90]. Therefore, the saturation constant was set to 0.01 and the constant '*a*' estimated, Table 4.2.

Table 4.2 Reaction rate constants determined via non-linear regression and saturation constants set to 0.01

Metabolite	a
Glucose	0.1929
Glutamine	0.0393
Lactate	0.2918
Ammonia	0.0421
Alanine	0.0184

After applying the reaction rate constants to a dynamic simulation a simulated fit can be overlayed the sampled data from Bastin and Provost's work, Figure 4.10 [90]. The fit achieved is a good prediction of concentration for both substrates and products. Fit values for Figure 4.10 have been determined, along with fit values for the dynamic simulation presented in Bastin *et al*'s work, Table 4.3.

Table 4.3 Fit achieved for each extracellular metabolite vs time plot using Provost and Bastin et al's simulation and the iMFA generated simulation

Simulation	Biomass	Glucose	Glutamine	Lactate	Ammonia	Alanine
iMFA	0.3654	0.2455	0.7078	0.4674	0.2633	0.5606
Bastin et al	0.3654	0.2756	0.5823	0.3189	0.2595	0.2448

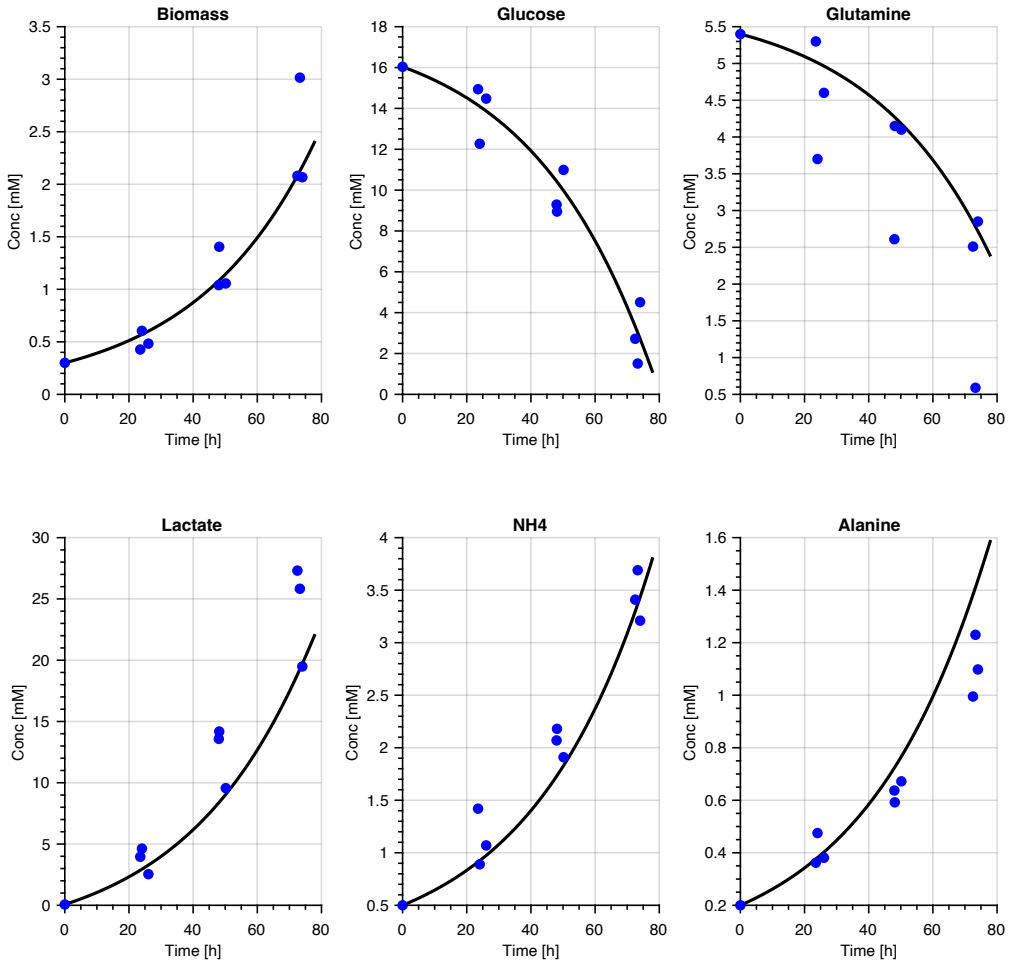

Figure 4.10 Sampled concentrations from Provost and Bastin experimental data (blue) and iMFA generated simulated model (black)

Bastin *et al*'s fit based on MFA is on average better performing than the one created with the iMFA results. iMFA is has a NRSME fit 31.28% worse than Bastin *et al*'s across the 6 metabolites, Table 4.3. However, the iMFA route offers a simpler method to create dynamic simulation than Bastin's. iMFA does not required measured flux rates to be known/or estimated; a process that can be lengthy and error introducing. Moreover, the proposed method only requires the stochiometric matrix and sample concentrations to be known. Estimation of biomass specific growth rate is still required. Overall, the iMFA method offers a reliable route to create a dynamic simulation of cell growth which requires less approximated data compared to traditional methods.

4.3.4 System Types

All results thus far have been on batch systems. To investigate the effects of operating under different conditions, both fed batch and continuous systems have been modelled with the hypothetical cell presented in section 4.2.2. Equations (4.33) to (4.35) detail the extracellular derivative equations used for modelling substrates, products and intracellular species conditions respectively [97]. These are substrate, product, and biomass specific versions of equation (4.6). D_i is the dilution rate, x biomass concentration and r reaction rate. The equations adjust based upon the system type. For a batch system, $D_i = 0$ and for a fed-batch system $[P] = [P^f]$.

$$\frac{dc_S}{dt} = D_i(c_{F,S} - c_S) - r_j X \tag{4.33}$$

$$\frac{dc_P}{dt} = r_j X - D_i c_P \tag{4.34}$$

$$\frac{dX_{macro}}{dt} = r_{macro} - \mu X_{macro} \tag{4.35}$$

Figure 4.11 and Figure 4.12 provide the extracellular metabolite concentrations over the simulation time for continuous and fed batch systems respectively. A typical continuous bioreactor is the chemosat, where the medium is designed to ensure there is only one single rate-limiting substrate [106]. The growth rate of biomass can be controlled and often steady state is achieved in these systems [97]. Figure 4.11 shows an ideal system where the substrates 'A', 'G' and 'F' have constant concentrations in the vessel as the feed rate is equal to the consumption rate. The substrates have a period of initial growth and then their concentrations are constant as the extraction of these metabolites form the vessel is equivalent to the accumulation. The iMFA results therefore can produce realistic models for continuous systems that mirror that of which is expected experimentally.

Continuous systems are not often used due to the risk of contamination from the feed stream [97]. Instead fed batch is the system of choice for many fermentations as the substrate feed can be kept constant [97]. Figure 4.12 shows the fed batch case where the inlet flow of substrates is less than the consumption rate (rates remained unchanged). It was therefore expected that 'A', 'G' and 'F' would decrease in concentration over time, which the iMFA results clearly show to be true. By modelling a case where the substrate consumption rate is greater than the inlet rate it allowed for checks to be made that the product concentrations responded by increasing and then plateauing as the substate concentrations dwindled. This is seen in the results of 'H' where the concentration begins to level off at ~100hrs.

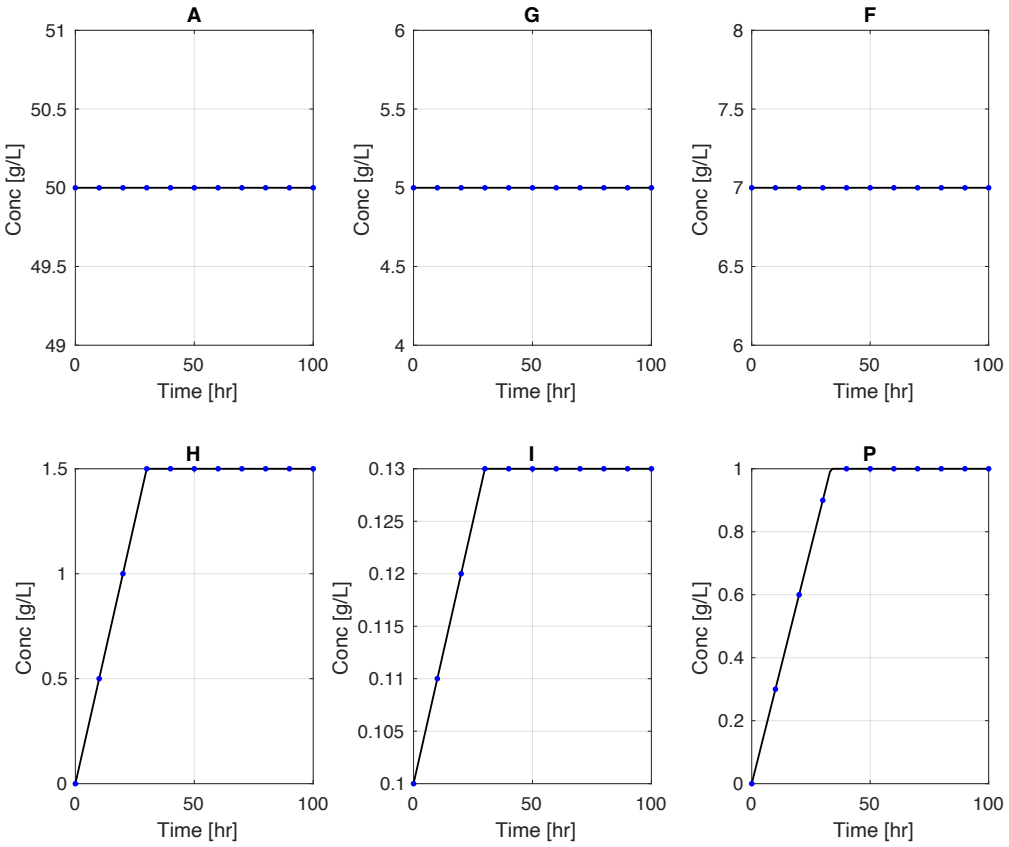

Figure 4.11 Extracellular concentrations A,G,F,H,I and P for a continuous system, blue dots showing sampling of 10 hours

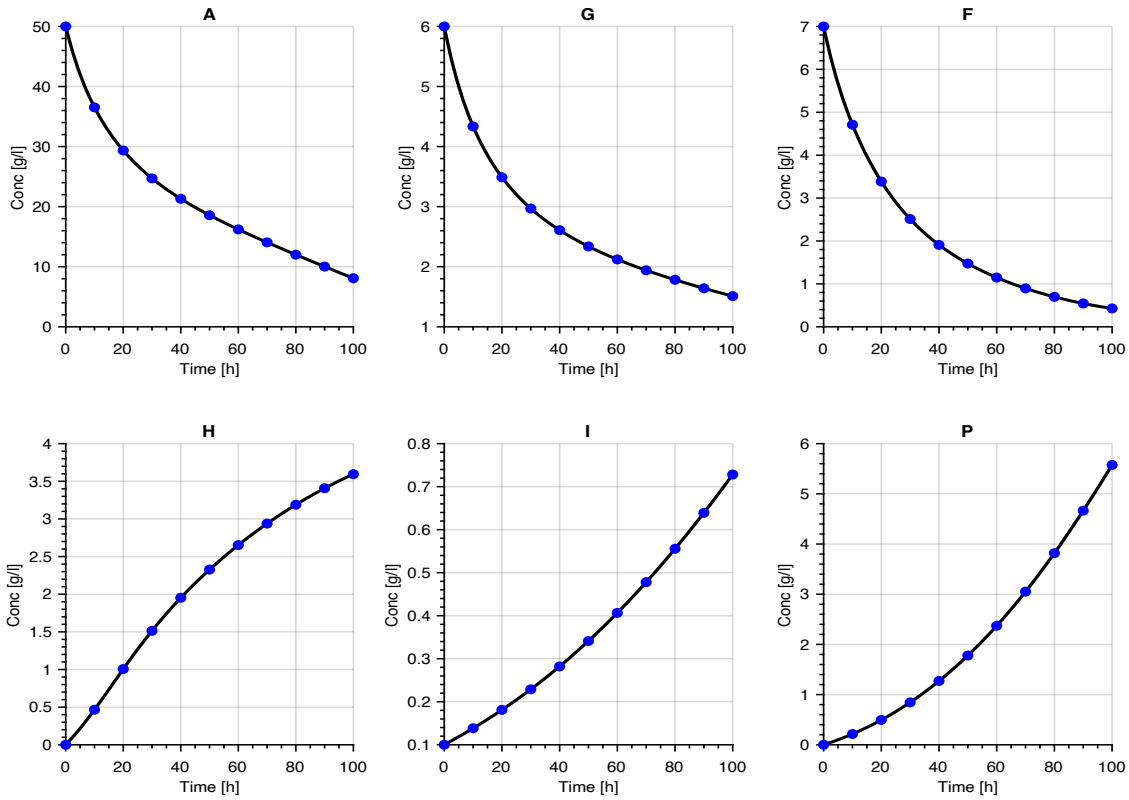

Figure 4.12 Extracellular concentrations, A,G,F,H,I and P for a fed-batch system, blue dots showing sampling of 10 hours

MFA and iMFA for the continuous system yield Figure 4.14 and fed batch Figure 4.15. iMFA appears to perform worse than MFA, however, the NRSME results reveal that iMFA is in fact more accurate, **Error! Reference source not found.**. MFA is better suited to batch systems over continuous and fed batch systems, **Error! Reference source not found.**. MFA relies on a derivative estimation, which for continuous and fed batch systems also include a dilution rate and current concentration values. Without ideal data, as in the simulation, the derivative approximation for these systems will have greater error than with the batch system.

iMFA outperforms MFA in all systems. The good fit values across the board show that it should be the preferable choice over MFA. The fed batch and continuous systems offer a poorer fit than batch with the iMFA. Like with MFA this is due to the introduction of error form noisy concentration data. However, iMFA provides on average 2.24 times better fit than MFA, **Error! Reference source not found.**.

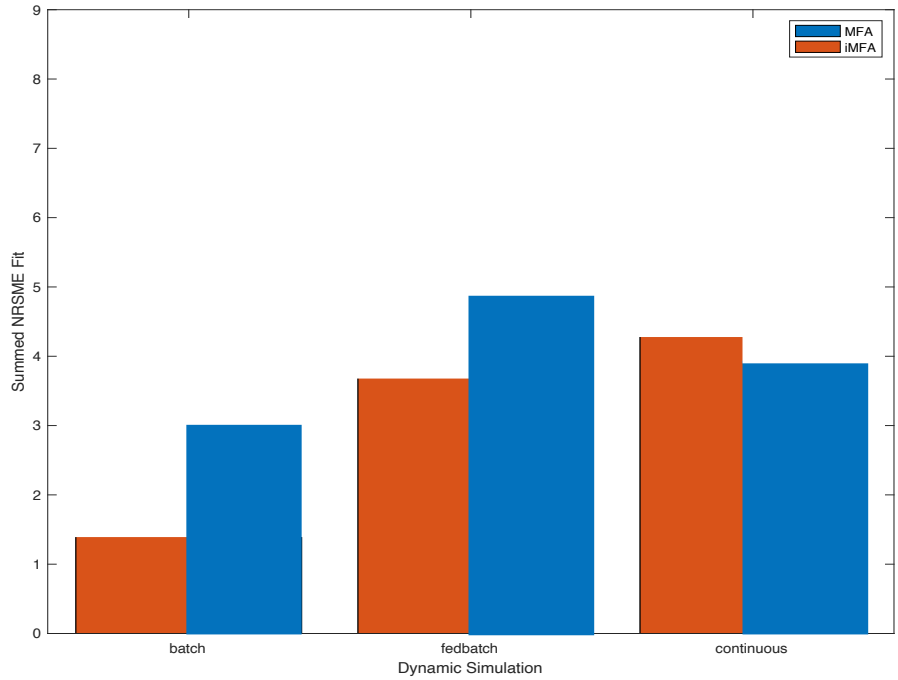

Figure 4.13 Dynamic simulation summed fits achieved for MFA and iMFA

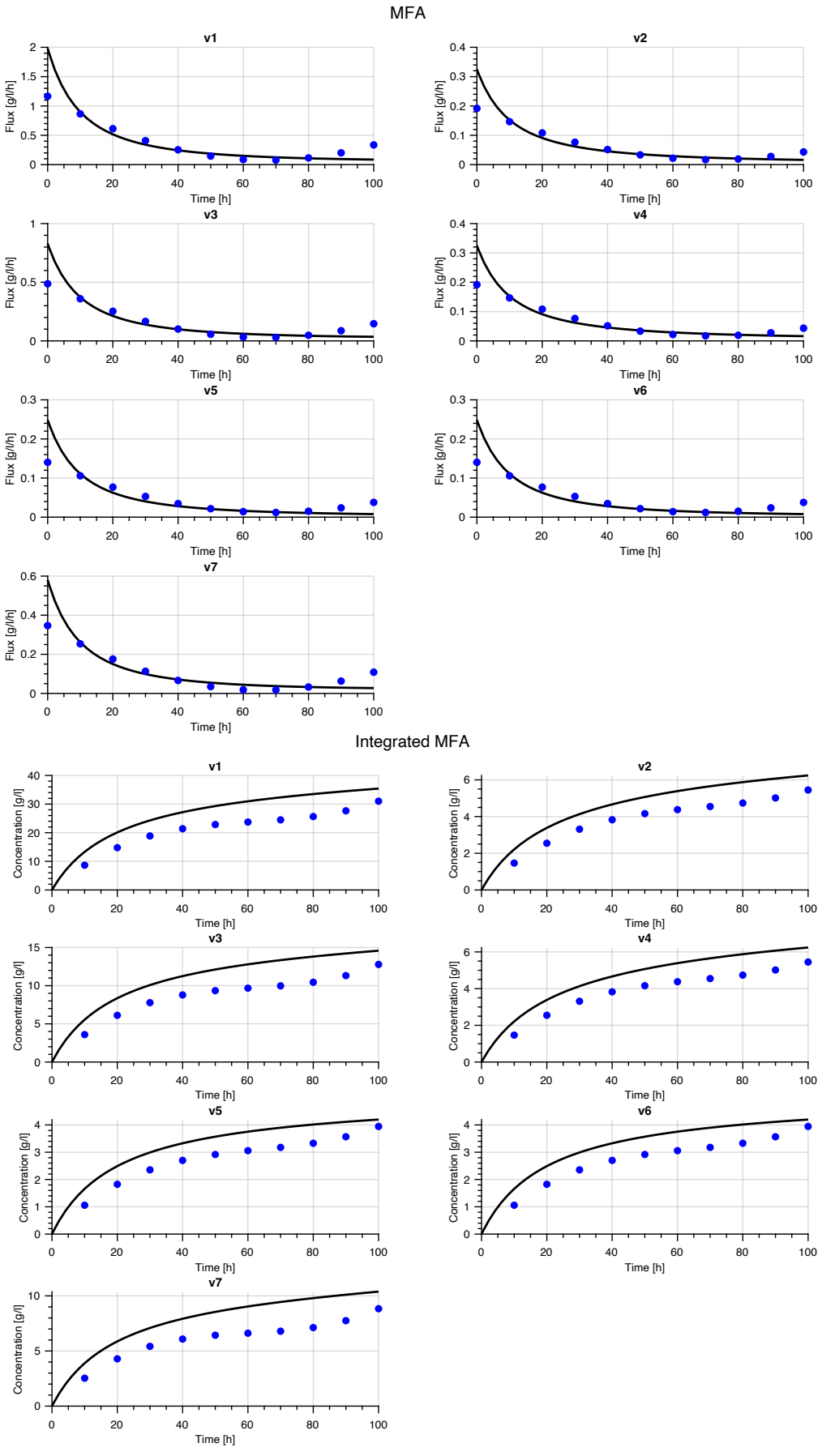

Figure 4.14 Continuous system with MFA and iMFA. Ideal data (black) and approximated with polyfitting (blue dots)

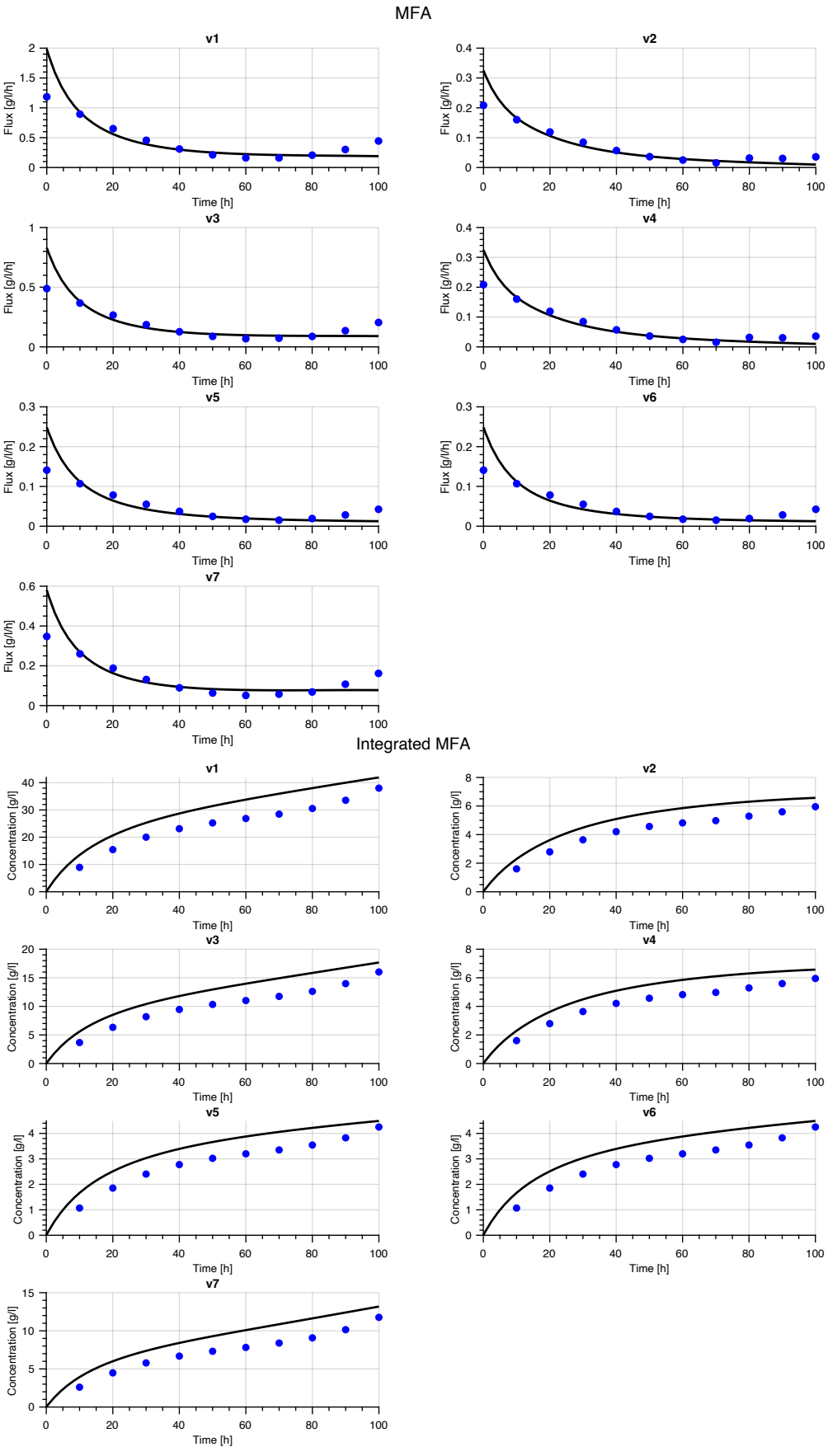

Figure 4.15 Fed batch system with MFA and iMFA. Ideal data (black) and approximated with polyfitting (blue dots)

4.3.5 Conservation Relationships

Results from iMFA can be utilised to define the conservation relationships of any cell network. These relationships relate substates to products. Every solution or operating mode is contained within the null space of the stochiometric matrix [107]. If we wish to find the macroscopic conservation relationships; where mass consumed is equivalent to mass produced [108], then the null space of the overall stoichiometry must be found, equation (4.36).

$$\boldsymbol{m.b} = null\left(\boldsymbol{S_K}^T\right) = \begin{bmatrix} 0 & 0 & 0.5 \\ 0 & 1 & -1 \\ 1 & 0 & -0.5 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{4.36}$$

In this equation $\boldsymbol{m.b}$ $(m_{ex} \times e)$ is matrix of conservation relationships. The overall stoichiometry is used as we are only interested in the minimal routes that give a macroscopic view of the network. The results of the iMFA can be used to expand upon this showing how changes in substrate effect product yield. The balances for this system created using the results of equation (4.36) are,

$$m_F = m_H \tag{4.37}$$

$$m_G = m_I \tag{4.38}$$

$$0.5(m_A - m_F) - m_G -= m_P \tag{4.39}$$

Therefore, the consumption of F is directly proportional to the production of H, so on so forth. To expand on these balances a MILP method can be applied to minimise the sum of the binaries associated with each metabolite within $\boldsymbol{S_K}$, equation (4.40). The equality constraint ensures the overall stochiometric matrix, multiplied by some vector $\boldsymbol{c}$ $(1 \times \boldsymbol{m})$, is equal to 0, equation (4.41). Equation (4.42) shows that δ_i is the binary variable vector associated with any extracellular metabolite. This, combined with integer cuts, equation (4.43), will provide all the possible conservation relationships that exist within the network. In equation (4.43) $\boldsymbol{\delta}_a$

are the vectors of binary variables obtained from the previous k iterations of the MILP and $\|\boldsymbol{\delta}_a\|_0$ is the cardinality of the vector of binary variables obtained at iteration k.

$$min \sum_{i=1}^{n} \boldsymbol{\delta_i} \tag{4.40}$$

$$\boldsymbol{S_K} \cdot \boldsymbol{c} = \boldsymbol{0} \tag{4.41}$$

$$\boldsymbol{\delta_i} = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_n \end{bmatrix}, if\ c \geq 1, \delta_n = 1\ else\ \delta_n = 0 \tag{4.42}$$

$$(2\boldsymbol{\delta}_a - 1)^T \boldsymbol{\delta}_{k+1} \leq \|\boldsymbol{\delta}_a\|_0 - 1 \quad (a = 2, \dots, k) \tag{4.43}$$

The application of this method finds two of the already defined conservation relationships, equations (4.37) to (4.39), along with the following.

$$0.49m_G + 0.01m_F = 0.49m_I + 0.01m_H \tag{4.44}$$

$$0.167m_A = 0.33m_I + 0.167m_H + 0.33m_P \tag{4.45}$$

$$0.01m_A + 0.47m_F = 0.02m_I + 0.48m_H + 0.02m_P \tag{4.46}$$

$$0.01m_A + 0.47m_G = 0.49m_I + 0.01m_H + 0.02m_P \tag{4.47}$$

$$0.01m_A + 0.01m_G + 0.46m_F = 0.03m_I + 0.47m_H + 0.02m_P \tag{4.48}$$

These relationships inform us on the ratio of substrates to desired products. It also highlights the importance of one substrate over another for the network. For example, only small quantities of substrate '*A*' are required for all conservation relationships. If the culture was previously being grown with large amounts of all substrates, having this information would allow for substrate quantities to be decreased whilst not sacrificing product yield.

4.4 Conclusion

This chapter has focussed on the use of an integrated form of MFA to approximate material transfer through all routes of a metabolic network. This method uses the integrated form of flux – concentration change – to calculate intracellular data. A prevalent issue with MFA is it's

inaccuracy with real time data [94], however, iMFA has been found to be more accurate in these cases. iMFA's accuracy with sparse, noisy data sets means it is a viable tool for industry to monitor the transfer of material through routes in a network. Monitoring material transfer is a useful technique in understanding the phase of cell life. Also, if the transfer of material through routes are known, it would be possible to expand the work into examining which routes are active over the cell's lifetime.

iMFA was also found to be useful in creating a dynamic simulation of the cell's growth phase. Approximation of rate constants via lengthy methods of trial and error have been widely used [79, 90, 109]. Although successful, as in the case discussed in this work, sometimes a better fit than iMFA, the process is sped up via the use of iMFA. iMFA offers an efficient method in approximating both the saturation constant and rate constant for a reaction rate in a Michaelis-Menten form. Whilst still offering a good prediction of substrate and product concentrations over time. To add to this, iMFA does not require behaviour assumptions of the cell to be made, apart from the pseudo-steady state assumption. Specific fluxes are not necessary in the prediction of intracellular material change, a major drawback of MFA.

Derivative approximation is required to predict measured fluxes. For MFA this leads to the introduction of errors. The same can be said for iMFA but the errors are greatly reduced, with data reconciliation offering the best approximation for iMFA. By using the integrated form of derivative approximation for measured rates this work found that data reconciliation performed well across all sampling times. This is due to data reconciliation reducing the errors on the individual data points rather than using the noisy data to estimate other values for use in the simulation.

The proposed method does not improve on MFA's inability to work on underdetermined systems or reversible reaction networks. However, the reduction in error in the solution achieved highlights the potential of using iMFA. FBA allows for the use of underdetermined systems; further to this it may be possible to use an integrated form, to build a predictive model of a metabolic network using material change. Therefore, it may also be possible to use an integrated form of FVA.

Using metabolic analysis methods is widely done, with improvements often only focussing on wholly new techniques. The adjustment of current techniques may offer an avenue of research to understand a cell's network in the manner of material change instead of flux. The reduction of error in real time data across batch, fed batch and continuous systems make iMFA a real contender in flux analysis techniques.

Chapter 5 Elementary Flux Modes

5.1 Introduction

An elementary flux mode (EFM) is a non-decomposable route through a cell connecting an extracellular substrate, through an intracellular network, to an extracellular product. A set of EFMs describe all the routes through a network. The number of EFMs in a network increases with the number of reactions present, for example a medium sized network for Escherichia coli (E.coli) (containing about 100 reactions) consists of around 272 million EFMs [25]. The main issue associated with this computation is the combinatorial explosion of the number of EFMs with the network size [87], additionally requiring a large memory space.

This chapter discusses a mixed integer linear programming (MILP) approach to solve EFMs and how it compares to publicly available tools; efmtool and FluxModeCalculator [29, 30, 31, 33]. The developed approach is an extension to the algorithm first proposed by de Figueiredo *et al* originally designed to calculate the K-shortest EFM's, where K is any assigned integer determining how many EFMs you wish to find [32]. The approach includes additional model constraints that enhance the efficiency of the MILP algorithm which allows larger networks to be analysed. In de Figueiredo *et al's* work only the 10-shortest EFMs were ever found for genome size networks. There also was no report on the speed of the method in finding EFMs within networks of increasing complexity. The number of EFMs solvable will be discussed along with how MILP could be used in the future to solve large scale networks. Heuristics, an approach to solving the problem using biological fundamentals have also been employed to reduce the EFM search space to ensure only EFMs that can occur will be found.

The use of linear programming techniques, particularly mixed integer linear programming (MILP) to determine EFMs has also been examined throughout literature. In work by de Figueiredo *et al*, EFMs are enumerated via a sequence of MILP optimisation problems [32]. Any flux mode with the minimum number of reactions must be, by definition, an EFM. This acts as the basis of the method which is used to find the 'K-shortest' EFMs for a particular metabolite production or consumption. Although this technique is successful, the authors acknowledged that it is better suited for product/substrate targeted EFM enumeration. This is

the best way to get biologically significant and therefore meaningful results from EFM enumeration.

Pey and Planes [110] proposed using MILP to enumerate EFMs that fulfil several biological constraints. Along with constraints from de Figueredo *et al*'s work, the use of the additional biological constraints reduced problem size and enhanced the MILP solver solution time. This was applied to a saccharomyces cerevisiae network to maximise ethanol production via the use of 100% of the carbon source, glucose. To do this they deactivated reactions which did not contribute to ethanol's production to get the subset of EFMs.

Larger scale EFM enumeration has been possible with MILP as the basis. Chan *et al* proposed an MILP scheme to break down flux distributions into EFMs at genome scale [111]. Unlike the other methods, this algorithm finds a set of EFMs that decompose a flux distribution for large scale networks, without the prior need for all EFMs. This work could go further still by examining several flux vectors simultaneously to reveal common routes across experimental conditions. However, it provides an analytical method ready for use if *in vivo* flux measurements can be performed.

MILP has also been used in other manners to examine a metabolic network. For example, Kaleta *et al* proposed using it to find pathways in sub-networks, within the context of the whole network [112]. These pathways are known as elementary flux patterns (EFP). EFPs can correspond to at least one EFM in a network; so are useful in determining the robustness of the network and the composition of minimal substrates need for the production of a particular product [59]. Another example, is the work of Bockmayr and Röhl which reduced networks down into subnetworks with given properties [113]. EFM enumeration in smaller networks is less computationally hard so splitting into subsections allows for easier solving. These works show the multiple uses for MILP on metabolic networks, outside the solving of EFMs.

This chapter proposes the use of MILP as a successful method for the determination of EFMs. Established commercial solvers use variations of the double descriptive method only and the proposed method offers a reliable technique ready for computational advances in the following years. Any future advances in hardware and solver efficiency will improve the solve

time of MILP. In 24 years, up to 2014, there was a 200 billion speed up in the solving of MILP problems, with further advancements being made every year [114, 115, 116].

5.3 Methods

A flux distribution within a metabolic network is any vector of flux rates $\boldsymbol{v}$ $(N_r\ x\ 1)$ that satisfies equation (5.1), where $\boldsymbol{S}$ $(N_m \text{ x } N_r)$ is the internal stoichiometric matrix comprising of N_m metabolites and N_r reactions. To ensure that only irreversible reactions exist, any reversible reaction is decomposed into two irreversible reactions (as discussed in chapter 2). Furthermore, the non-decomposability of a EFM is represented by equation (5.2), which states that, an EFM is a minimal, unique set of flux-carrying reactions.

$$\boldsymbol{v} \in \{\boldsymbol{v}|\boldsymbol{Sv} = \boldsymbol{0}, \boldsymbol{v_{irr}} \geq \boldsymbol{0}\} \tag{5.1}$$

$$\boldsymbol{E_1} \not\subset \boldsymbol{E_2} \tag{5.2}$$

MILP can be used to find EFMs due to the constraint-based nature of the problem. For each reaction in the metabolic network a binary integer variable, $\delta_i (i = 1, \ldots, N_R) \in \{0,1\}$ is used to define whether a reactions is active, $\delta_i = 1, \Rightarrow v_i > 0$ or not, $\delta_i = 0, \Longrightarrow v_i = 0$. The MILP approach proposed by de Figueiredo *et al* aimed to minimise the sum of these variables subject to the additional constraints discussed below [32]. The decision variables are the on/off binary variables representing reaction pathways.

$$\min \textstyle\sum_{i=1}^{N_R} \delta_i \tag{5.3}$$

$$\boldsymbol{Sv} = \boldsymbol{0} \tag{5.4}$$

$$v_i \leq M_i \delta_i \qquad (i = 1, \ldots, N_r) \tag{5.5}$$

$$\delta_i \leq v_i \qquad (i = 1, \ldots, N_r) \tag{5.6}$$

$$\sum_{i=1}^{N_r} \delta_{\boldsymbol{i}} \geq 1 \tag{5.7}$$

Constraint (5.4) ensures the steady-state condition of the metabolic network and determination of the flux distribution associated with the active reactions. Equations (5.5) and (5.6) represent the so-called Big M constraint method, where M is the Big M constant (which

theoretically can be different for each reaction). These equations ensure that $\delta_i = 1, \Rightarrow v_i > 0$ and, $\delta_i = 0, \Longrightarrow v_i = 0$. The choice of the Big M constant is always greater than the greatest stochiometric ratio allowed within any flux distribution [111]. In addition, equation (5.7) ensures that at least one reaction is active avoiding the trivial solution where all the binary variables are set to zero.

A final constraint is added to the MILP to ensure any reactions from the same reversible reaction set cannot simultaneously occur; recall, reversible reactions must be decomposed into 2 individual reactions to ensure equation (5.7) is met. Assuming the set of reversible reaction pairs is defined, then for each reversible reaction, there will be a binary variable associated with the forward reaction, $\delta_{f,l}$ and the reverse reaction $\delta_{r,l} \quad (l = 1, \dots, N_{rev})$ where N_{rev} are the total number of reversible reactions in the network. Therefore constraint (5.8) ensures that the reversible reaction pair cannot occur simultaneously.

$$\delta_{f,l} + \delta_{r,l} \leq 1 \qquad (l = 1, \dots., N_{rev}) \tag{5.8}$$

The solution of this MILP will find one potential EFM, that must be confirmed. This additional test is discussed in section 5.3.2. Furthermore, as formulated, equation (5.3) to (5.8) will only find one solution (the shortest – in terms of the total number of reactions included) and to find all solutions (all EFM's) integer cuts are required so that once a potential EFM is found it is removed from the solution space and the algorithm is resolved to find additional EFMs.

5.2.1 Integer cuts

When the solution to the MILP is found an integer cut is added, equation (5.9) [117], before the solution procedure is repeated for $k = 1, \dots.. N_k$ iterations.

$$(2\boldsymbol{\delta}_a - 1)^T \boldsymbol{\delta}_{k+1} \leq \|\boldsymbol{\delta}_a\|_0 - 1 \qquad (a = 2, \dots, k) \tag{5.9}$$

In equation (5.9) $\boldsymbol{\delta}_a$ are the vectors of binary variables obtained from the previous k iterations of the MILP and $\|\boldsymbol{\delta}_a\|_0$ is the cardinality of the vector of binary variables obtained at iteration k.

5.3.2 The rank test

To check the solution is an EFM, a rank test is used. This has been extensively used throughout literature [27, 31, 118]. If all reactions are irreversible, then $\boldsymbol{v}$ is an EFM if equation (5.10) is true. This test works as a feasible solution is an EFM if and only if the null space of the submatrix $\boldsymbol{S_{EFM}}$ that involves the reactions of the EFM is of the dimension one [27]. In other words, the rank of this submatrix must be equal to the number of the participating reactions minus one.

$$\rho(\boldsymbol{S_{EFM}}) = \|\boldsymbol{\delta}_a\|_0 - 1 \tag{5.10}$$

Note – this test is done after all iterations have been performed by the MILP. Figure 5.1 provides a flowchart view of the proposed method prior to the addition of further constraints.

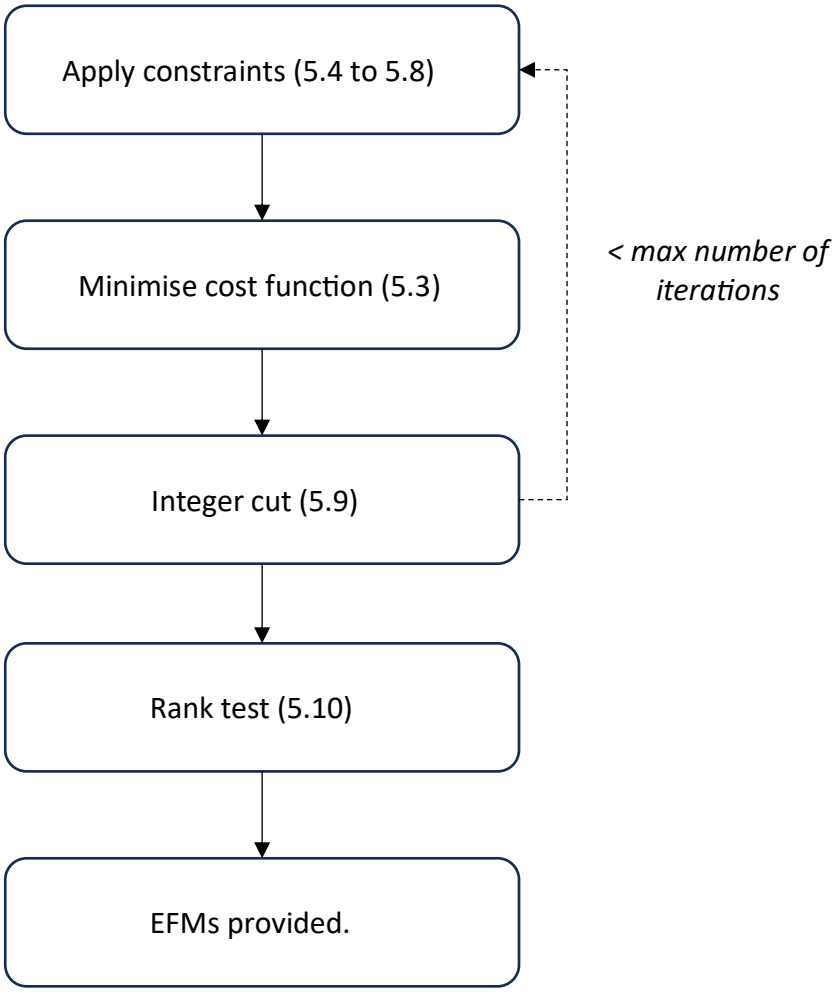

Figure 5.1 Flow chart of EFM enumeration via proposed MILP method

5.3.3 Additional constraints to improve the efficiency of the MILP

There will be a binary variable associated with the reactions directly related to the substrates, $\delta_{s,m} \quad (m = 1, \dots, N_s)$ and the products $\delta_{p,n} \quad (n = 1, \dots, N_p)$ where N_s is the total number of substrates and N_p the total number of products. Therefore, to ensure that at least one substrate and at least one product is included in any EFM the following constraints can be defined,

$$\sum_{m=1}^{N_s} \delta_{\boldsymbol{s,m}} \geq 1 \tag{5.11}$$

$$\sum_{n=1}^{N_p} \delta_{\boldsymbol{p,n}} \geq 1 \tag{5.12}$$

Any EFM must be a minimal decomposition of the flux distribution and cannot share the same reaction set with another EFM, equation (5.1). This definition helps in the reduction of the search space as once a feasible EFM has been found no other EFM can contain the exact set of reactions as another. Operating similarly to an integer cut, constraint (5.13) ensures that the new EFM can be composed of at most $\|\boldsymbol{\delta_k}\|_0 - 1$ of the reactions in the previously found EFMs.

$$\boldsymbol{\delta_a^T \delta_{k+1}} \leq \|\boldsymbol{\delta_a}\|_0 - 1 \quad (a = 2, \dots, k) \tag{5.13}$$

Dependency of one reaction onto another can be observed through flux balances. If any metabolite is consumed by only one reaction, $\delta_{c,1}$, and produced by only one reaction, $\delta_{pr,1}$ then constraint (5.14) must hold. This is commonly known as flux coupling and is often used to reduce the size of genome networks [119, 120].

$$\delta_{c,1} - \delta_{pr,1} = 0 \tag{5.14}$$

Constraint (5.14) effectively compresses the network, reducing the search space. These flux balance constraints can be deduced from a *priori* inspection of the rows of the stoichiometric matrix $\boldsymbol{S}$. Alternatively, they are observed through the digraph representation of the network whereby species nodes that an indegree and an outdegree of one identify the reaction pairings.

5.3 Algorithm implementation

The MILP is implemented in MATLAB using the '*intlinprog'* function. We make use of the output function associated with '*intlinprog'* as this reports additional solutions to the MILP at each iteration. Not all these solutions will be EFMs when tested, however, by saving all outputs as possible solutions the search time may be reduced as some will be EFMs (the benefit of using this will be reported in the results section).

Maximum iterations are manually set for all networks. These are particularly important when working with larger, complex networks as the MILP branch and bound search method will take time to find feasible solutions.

5.4 Results

5.4.1 EFM Detection

A total of 9 networks were initially tested in the MILP set-up for finding EFMs. Four MILP setups were trailed: 1) standard MILP equations (5.8) to (5.10), 2) use of the *intlinprog'* output function which reports additional (sub-optimal) solutions to the current MILP, 3) constraint (5.13) preventing reaction sets reappearing in future EFMs and 4) constraint (5.14) to compress the network through flux balances. All these tests have been timed and compared to the results found in efmtool and the FluxModeCalculator (FMC). Each network was chosen or designed to test that the solver was able to find EFMs with varying circumstance. The networks tested are as follows (their sizes, i.e., number of metabolites, number of reactions and number of reversible reactions can be found in Table 5.1):

1. A realistic simple network for initial EFM detection [93].
2. Knockout reaction network [70].
3. Larger realistic network – Chinese Hamster ovary (CHO) cell – which is exactly determinable [76].
4. Larger realistic network – CHO cell – which is underdetermined [76].
5. Simple saccharomyces cerevisiae network [46]
6. Pichia Pastoris cell– pyruvic acid acting as a product as per the stoichiometry used in literature [121].
7. Simple yeast core model based on that used by Damiani *et al*, [122].
8. Pichia Pastoris cell– pyruvic acid acting as a substrate [121].
9. Escherichia coli (*E. coli*) core [38].

Table 5.1 Network sizes, number of flux balances and the value of the BigM used in EFM enumeration

Network	No. Reactions (reversible pairs)	No. Metabolites	Flux balances	Big M
1	7 (0)	10	2	2.5
2	14 (0)	15	4	1
3	19 (0)	21	5	10
4	24 (0)	25	5	10
5	22 (0)	21	2	10
6	61 (17)	46	10	3500
7	40 (9)	26	5	10
8	61 (17)	46	9	781 < BigM < 45833
9	114 (39)	92	2	1000

The Big M value, equation (5.5), varies with each network. For the networks discussed in section 4.5.1 the Big M values are given in Table 5.1. Larger networks with more reversible reactions require larger Big M values. If the Big M value chosen is too small not all EFMs will be found. Often the Big M is chosen to be the upper bound on the flux range across the reactions [113]. For networks with integer stoichiometry, flux estimates are often small integer values that are scalar with the stoichiometry. However, network 8 consists of stoichiometry which are not integers, leading to a large range of flux estimations. These values are reflected in the Big M's chosen. There is a large range due to the number of reactions each with non-integer values. To improve the efficiency, it was decided to use differing Big Ms based upon the individual reactions upper flux estimations. Table 5.1 also details the number of flux balances per network used to reduce the search space, equation (5.14).

Table 5.2 gives all the computation times and efficiency for all 4 MILP methods discussed, efmtool and FMC. Efficiency is the number of true EFMs, confirmed via the rank test, equation (5.10), divided by the number of potential EFMs found by the solver. This solve time is computer specific; MacBook Pro; 2.4 GHz Quad-Core Intel Core i5, 8 GB 2133 MHz LPDDR3. Efmtool outperformed FMC's computation time throughout all tests, however, this was expected as it reported in literature that efmtool is the superior EFM solver [53, 54].

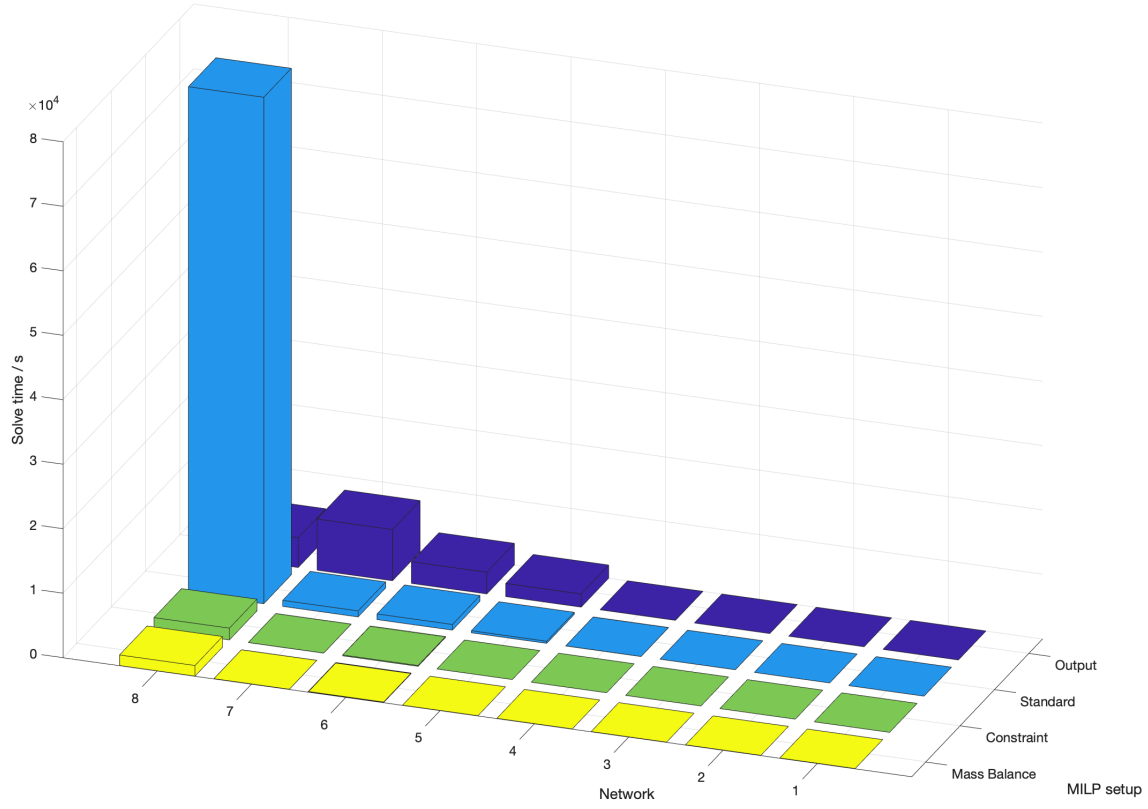

Figure 5.2 Run times for networks across all the MILP setup with network 8

Figure 5.2 provides the run times for all networks (excluding network 9 due to large network size and run times) across the various MILP setups. Network 8 in standard MILP form has a large run time, making it difficult to interpret the results. Therefore, Figure 5.3 also displays the run times, but without network 8. Network 8's run time is much larger than the others due to the increased number of possible EFMs which all require storing as the MILP iteratively solves for the next EFM. In the standard MILP form the branch and bounding is a time-consuming process. However, when other constraints are added and the output function is used, finding more solutions each iteration, the solve time is drastically reduced.

Figure 5.3 shows that the output function tends to take longer to run with larger networks than the standard MILP. This is due to increased memory storage being required, which in turn reduces the efficiency of the solver. It is clear, however, that the addition of the extra constraint (equation 4.14) and flux balances with the output function reduce solve time across all networks considerably.

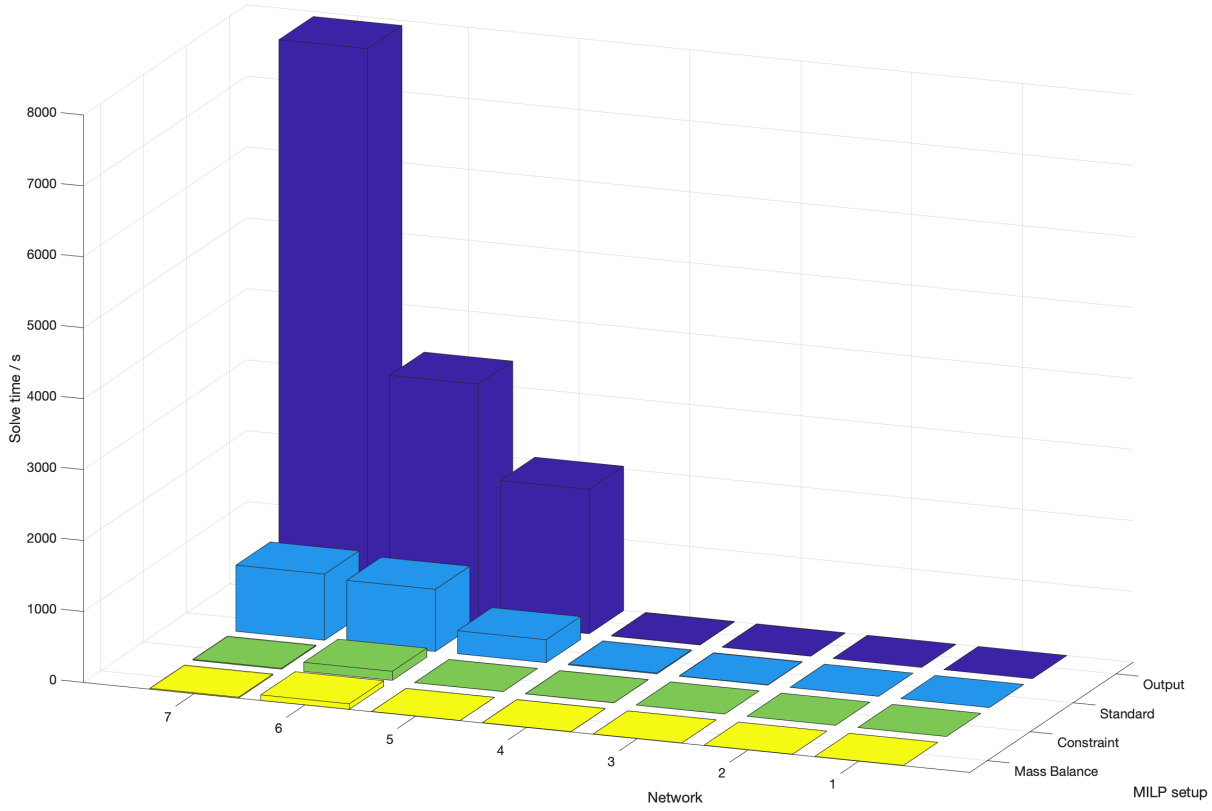

Figure 5.3 Run times for networks across all the MILP setups without network 8

Run times for network 9 were longer than 24 hours for the standard MILP so the run time for both these networks were limited to 10 minutes and 1 hour. The EFMs found in this time and the efficiency are given in Table 5.3. The number of iterations of the MILP was set to 100 with a target EFM of 500. Both efmtool and FMC could find all 100274 EFMs of the *E. coli* core in 37.38s and 2min 13.40s respectively. Efmtool's efficiency for the *E. coli* core is 0.9999 and FMC's is not reported as it is not generated by the solver. The best results for number of EFMs found in this case was when all constraints were applied with MILP. Efficiency also improves

when all constraints are used as the simulation time increases. Run times increase as the networks do, as the number of constraints generated rises.

Table 5.2 Run times and efficiency for various MILP methods, efmtool and FluxModeCalculator. No efficiency is provided for FluxModeCalculator this is unknown. Network 9 and 10 computation times for FMC and efmtool are for the full 100,274 EFMs not the reduced amounts set for MILP

Network Number	Variables	Standard MILP	Output Function	Constraint (5.13)	Flux balance (5.14)	Efmtool	FMC
1	*Computation Time*	0.38s	0.37s	0.42s	0.58s	0.04s	0.31s
	Efficiency	1	1	1	1	1	-
2	*Computation Time*	0.85s	0.72s	0.46s	0.51s	0.05s	0.35s
	Efficiency	0.4	0.4	1	1	1	-
3	*Computation Time*	4.06s	3.04s	1.01s	1.04s	0.07s	0.69s
	Efficiency	0.3103	0.3103	1	1	1	-
4	*Computation Time*	12.82s	5.33s	2.16s	1.96s	0.02s	0.68s
	Efficiency	0.1864	0.1864	1	0.9167	1	-
5	*Computation Time*	5min 23.84s	33min 55.40s	1.23s	1.37s	0.33s	1.34s

	Efficiency	0.0575	0.0131	0.9583	1	1	-
6	*Computation Time*	14min 35.70s	55min 58.80s	2min 8.15s	77.21s	0.12s	6.19s
	Efficiency	0.0967	0.1106	1	0.9899	0.9074	-
7	*Computation Time*	8min 32.42s	1hr 6min 18.54s	6.74s	6.72s	0.05s	4.71s
	Efficiency	0.0844	0.0781	0.971	0.9853	0.9178	-
8	*Computation Time*	21hr 4.00s	1hr 17min	30min 26.20s	27min 26.40s	0.08s	29.61s
	Efficiency	0.0955	0.0719	0.9026	0.9287	0.9698	-

Table 5.3 E. coli network efficiency and number of EFMs found in 10 minutes and 1 hour

Run Time	Variables	Standard MILP	Output Function	Constraint (5.13)	Flux balance (5.14)
10 min	*Number of EFMs found*	62	144	194	168
	Efficiency	0.7949	0.5199	0.7791	0.7636
1 hour	*Number of EFMs found*	38	175	271	337
	Efficiency	0.7917	0.4861	0.7901	0.8180

5.4.2 Heuristics

Heuristics can be used to reduce the search space for EFM's based upon biological knowns. This may not find all EFMs, but instead will find the most biologically significant ones and the ones that are most likely to occur.

Any cell requires a carbon source to grow. Carbon sources taken by the cell act as a substrate to the network, where they can be broken down to supply pools of amino acids and other components [123]. Glucose is commonly used and is present in every network simulated in section 4.5.1. Therefore, reactions that provide the cell with glucose, or any carbon source, can be assumed to be active when searching for EFMs with MILP.

Cells can be aerobic, anaerobic, or facultative anaerobic, i.e., can grow in both aerobic and anaerobic environments. *E. coli* is an example of a facultative anaerobic cell [124]. For facultative cell's the oxygen reactions entering the network can be on or off dependent on the growth conditions. This can be used for reducing the search space when finding EFMs with MILP as the reaction may or may not be possible.

Chemical energy generated by substrate oxidations is conserved in bacterial cells by the formation of high-energy compounds such as adenosine diphosphate (ADP) and adenosine triphosphate (ATP) [125]. These compounds are vital for cell life. Simplified cell networks, like

the *E. coli* core, often have ATP maintenance reactions [38]. If these are present the search space should be adjusted to include some EFMs that encapsulate this vital area. This will highlight key areas of the network that have to be in operation to ensure high energy yields.

The tricarboxylic acid (TCA) cycle operates during aerobic and anaerobic respiration or fermentation by running in an oxidative cycle (when respiring oxygen) or in an incomplete, reductive, and branched pathway, respectively [126]. The cycle does not produce high amounts of ATP; but instead removes electrons from inputs and transfers them to an electron carrier that deposits their electrons onto the electron transport chain [127]. It is the most important central route connecting almost all the individual metabolic network parts, and therefore, ensuring it is in the EFM set is valuable [128]. Acetyl-CoA (acetyl coenzyme A) is the precursor to the TCA cycle, and therefore, in good cell growth conditions the reactions producing and consuming it should be active. Modification of the TCA cycle yields the glyoxylate cycle. Its main purpose is anabolic, to allow for the production of glucose from fatty acids. It is therefore essential for carbon sources such as acetate or fatty acids [129].

Any combination of basic biological constraints can be added to the MILP solver to reduce the search space to find fewer EFMs. This not only reduces the search time but highlights the biologically significant EFMs. It is important to note that applying all these constraints at once will only provide the EFMs that contain all the desired reactions, therefore, if the desire was to know individual reactions, then these must be computed individually.

For example, for the simplified CHO cell during the growth phase, glucose consumption will be high. Therefore, to reduce the search space it would be worth only permitting EFMs with reactions that consume glucose. As a cell is growing, CO_2 will be produced as the oxygen uptake increases if the cell is grown aerobically. Therefore, during the growth phase it also is worth including CO_2 production reactions. However, for a CHO cell it has been found that CO_2 levels of 36 to 250mmHG reduces the specific growth rate [130]. Therefore, although heuristically it makes sense to encapsulate CO_2 production during growth, the results may show EFMs that could in fact negatively affect the cell's growth. This will be discussed in further detail in Chapter 6. To simulate growth conditions, and therefore the only EFMs that could be occurring at that moment in time, in the EFM search space, reaction 1 for glucose

consumption and reactions 3,8,10,11 and 14 must be 'on' in the MILP solver (see Table 5.4 for the full reaction list). This reduces the number of EFMs from 9 to 2 and solve time from just over 1 minute to 0.18s. Table 5.4 shows the two EFMs found. If the reaction is denoted with a 0 for the EFM, the reaction is not required for the complete route of the EFM. The numerical values indicate the amount of ratio of the reaction required for the EFM, which is reflected in the fluxes obtained by flux analysis. In both the EFMs reactions 1,3,8,11 and 14 are present as this was specified in the MILP solver. The EFMs are very similar apart from $\boldsymbol{E_1}$ requiring reaction 7 and $\boldsymbol{E_2}$ requiring reaction 16. Both these reactions are ways to utilise glutamate in the production of α-Ketoglutaric. This is no surprise as α-Ketoglutaric is necessary to facilitate reaction 11, which was required via heuristic determination.

Table 5.4 Two EFMs found via heuristic determination of CHO cell

Reaction number	Reaction	$\boldsymbol{E_1}$	$\boldsymbol{E_2}$
r_1	$Glucose \rightarrow Glucose\ 6\text{-}phosphate$	1	1
r_2	$Glucose\ 6\text{-}phosphate \rightarrow Dilhydroxy\text{-}AP + Glyceraldehyde\ 3\text{-}phosphate$	0	0
r_3	$Glucose\ 6\text{-}phosphate \rightarrow Ribose\ 5\text{-}phosphate + CO_2$	1	1
r_4	$Dilhydroxy\text{-}AP \rightarrow Glyceraldehyde\ 3\text{-}phosphate$	0	0
r_5	$Glyceraldehyde\ 3\text{-}phosphate \rightarrow Pyruvate$	0	0
r_6	$Pyruvate \rightarrow Lactate$	0	0
r_7	$Pyruvate + Glutamate \rightarrow \alpha\text{-}Ketoglutaric$	0.5	0
r_8	$Pyruvate \rightarrow Acetyl\text{-}CoA + CO_2$	0.5	1
r_9	$Acetyl\text{-}CoA + Oxoloacetate \rightarrow Citrate$	0.5	1
r_{10}	$Citrate \rightarrow \alpha\text{-}Ketoglutaric + CO_2$	0.5	1
r_{11}	$\alpha\text{-}Ketoglutaric \rightarrow Fumarate + CO_2$	2	2.5
r_{12}	$Fumarate \rightarrow Malate$	2.5	3
r_{13}	$Malate \rightarrow Oxoloacetate$	1.5	2
r_{14}	$Malate \rightarrow Pyruvate + CO_2$	1	1
r_{15}	$Glutamate + Oxoloacetate \rightarrow Aspartate$	1	1
r_{16}	$Glutamate \rightarrow \alpha\text{-}Ketoglutaric + NH_4$	0	0.5
r_{17}	$Glutamine \rightarrow Glutamate + NH_4$	0	0
r_{18}	$Glutamine + Ribose\ 5\text{-}phosphate + Asparate + CO_2 \rightarrow Fumarate + Nucleotides$	0.5	0.5
r_{19}	$CO_{2,extracellular} \rightarrow CO_2$	4.5	6

The same method can be applied to a simplified yeast core network [122]. Glucose consumption is necessary for growth so consumption reactions in the stoichiometry must occur in any EFM result. The glyoxylate cycle and TCA should also be functioning as it is a vital central pathway. Ensuring any reactions in this cycle in aerobic respiration are on will ensure the EFMs found encapsulate this network section. Application of these heuristics highlights three key reactions. These three reactions reduce the EFMs down from 67 to 14 and the solve time from 12.66s to 7.89s.

Using heuristics to reduce the EFM space not only speeds up solve time but prevents biologically unfeasible solutions being produced. Targeting specific EFMs to understand how to maximise product yield and reduce waste will be discussed later in this thesis.

5.4.3 Essential Reactions

When computing some or all EFMs essential reactions become apparent. These are reactions that occur in most or all EFMs, meaning they are vital to the cell's life; biomass and other metabolite production would not be possible without them. Figure 5.4 shows the reaction usage across the 67 EFMs for the simplified yeast core cell. Reactions 4, 5, 14 and 28 are not

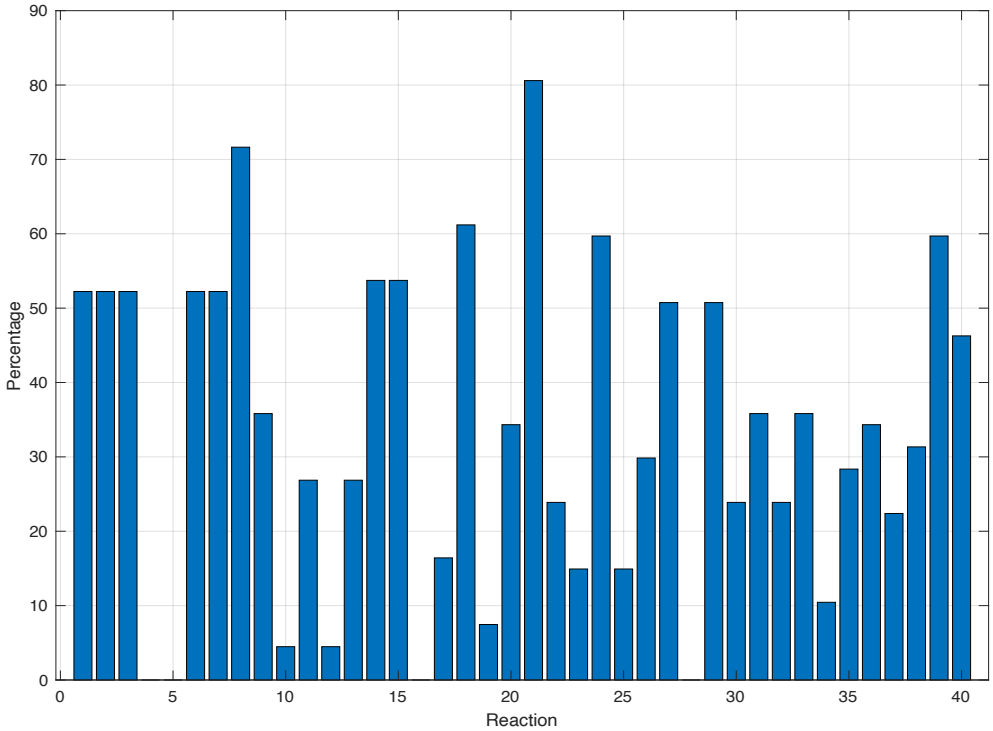

Figure 5.4 Reaction usage in EFMs for simple yeast core cell

used by any EFM. These are reactions that lead to dead-end metabolites, see section 2.8, and therefore removal of these from the search space would speed up solve time. The essential reactions are the ones with the greatest percentage, therefore, reactions 8 and 21 are the two most important reactions to the cell,

$$r_8 = Glucose \rightarrow Glucose\ 6\text{-}phosphate$$

$$r_{21} = Acetyl\text{-}CoA \rightarrow Malate + Glyoxylate$$

Reaction 8 is the consumption of intracellular glucose after it has been transported into the cell. Therefore, it is a precursor to most reactions within the network and is expected to be essential. This is reaffirmed with 72% of EFMs using it. Reaction 21 is the initialisation of the glyoxylate cycle, which as discussed in section 5.5.2, is vital to good cell growth. Overall, 81% of EFMs use reaction 21 This shows that even though there are no biological constraints in determining the full set of EFMs, the EFMs will highlight key areas required by the cell.

5.5 Discussion

5.5.1 Mixed Integer Linear Programming's General Performance

The MILP setup was unaffected by reversible reactions or an increase in network size, at this small scale. Despite FMC being published 7 years after efmtool, it often was 10 times slower than efmtool and at times was outperformed marginally by the MILP solver. Due to the efficiency of efmtool this solver will now solely be used for comparison with MILP throughout this thesis.

Overall, the MILP's performance is as expected – the efficiency reduces as the number of constraints and network size increase. The next steps were to test the limitations of MILP for this problem type. If the MILP solver was able to solve problems of a larger size, then it is just a matter of improving efficiency as it solves. It is positive to see similar operating times with the FMC in some examples as it shows that the current work was on track to eventually being on par with EFM solvers already publicly available. However, the key point to take away is that even though there MILP solves slower it has a greater chance of solving at genome scale due to how linear programming is solved via branch and bounding [111]. This is mainly due to the

ability to disregard constraints that are not relevant to the current search space due to their small size. The double descriptive method constantly builds over solve time and can only be minimised via sparsity or bit pattern trees without any disregarding.

Advances in solver efficiency and hardware also make MILP a promising option. Between 1990 and 2014 advances in integer optimisation, along with computational advances, led to a 200 billon factor speedup in solving MILP problems [114, 115, 116]. Bixby measured the speed up of MILP solvers by solving the same set of problems using twelve consecutive versions of CPLEX. The versions ranged from the 1991 release, CPLEX 1.2, up to the 2007 release, CPLEX 11. The total speed up factor, from the first to the final release tested, was 29000 [114]. Evidence of speed up in solvers is not just present in CPLEX, but also Gurobi. Gurobi 9.5, released 2021, is reportedly 15% faster on mixed integer problems than the previous version [131]. Hardware speed up between 1993 and 2013 is approximately $10^{5.5}$ [116]. This evidence all suggests the viability of MILP for solving EFMs at genome scale in the future.

The 'bench' test was performed in MATLAB to compare the device used to enumerate EFMs with other computer types. All other computer types outperformed the device used, and MATLAB online also offered considerable performance improvements.

Computing EFMs is a hard computational task, hence the restriction on the network size. The null space approach proposed by Wagner [58] offered improvement to the double descriptive method. It accelerated computation time and shifted most of the limitation over to the memory requirement for a typical PC [57]. Despite this development in 2004 it has still not been able to fully compute a genome scale network, with particular focus on activity the routes over time. The work thus far has shown the advantages of using MILP over the double descriptive method on network size, but now also offers an avenue to model the three phases of cell life alongside activity due to environmental factors; a feature currently not possible on widely used EFM solvers.

5.5.2 Impact of Network Features on Mixed Integer Linear Programming

The two main structural features of genome scale networks are reversible reactions and cyclic reaction sets. From the examples above it can be assumed that the cyclic system does not have a great impact on solve time. However, reversible reactions clearly do. In the MILP code produced, the presence of reversible reactions requires additional constraints to be added.

Another key part of a cell's metabolic network is the growth, transition, and decay phases. The existence of these three phases reduces the number of active pathways at any one time. This in turn reduces the number of EFMs to be calculated per phase. The utilisation of these three stages has been mentioned in literature already, however, the lack of true data accessible has meant that the use of these stages has yet to be used at genome scale [79, 93, 132]. Accurate application of this data would enable active sets of EFMs to be found easily within the MILP setup.

As a network increases in size so does the number of constraints. This slows the computation time due to the memory required to store the large matrices. After running the *E. coli* network for 10 minutes the inequality constraint matrix is 1.3MB. After an hour this matrix is 2MB in size. All matrices in the *E. coli* MILP setup, after an hour, total 5.94MB in size. Therefore, memory issues will become apparent as the networks increase in size and more EFMs are found. This is an ongoing issue with EFM enumeration and is not specific to MILP. However, parallelisation or high-performance computing (HPC) could offer a memory reduction with the utilisation of multiple cores and servers to store data.

5.6 Conclusion

This chapter presented a MILP approach with solve times for small networks and the *E. coli* core. The solve time of efmtool was quicker than MILP due to the many improvements made over the years to reduce memory usage and accelerate computation time [31]. However, improvements in solver efficiency and hardware present the opportunity for MILP to be a feasible solution to solving the problem in the future [114, 116]. To improve the solve time the next chapter will examine compression methods, to reduce solve time, and techniques to reduce memory storage requirements.

Chapter 6 Improving Elementary Flux Mode Discovery whilst using MILP

6.1 Introduction

Large metabolic networks require extensive memory storage and computational power to determine elementary flux modes (EFMs). This is the case for commercial EFM solvers, like efmtool, and the mixed integer linear programming (MILP) method presented in Chapter 5. To overcome the demands set by this large-scale network, reduction of its size has been exploited to improve performance [35, 113, 133, 134, 135, 136].

In 2015, Erdrich *et al*, introduced a method called NetworkReducer [35]. This method reduces large networks into smaller subnetworks whilst ensuring important biological properties remain, such as energy maintenance. NetworkReducer consists of a pruning and a compression step. In the compression step any reactions belonging to the same enzyme subset are lumped together. An enzyme subset is defined as a group of enzymes that operate together in fixed flux proportions in all steady states of the system [51]. The method searches for a suitable subnetwork by iterating over the reactions. The flux rate for one reaction is set to zero for each iteration and a linear program is solved to check if the remaining reactions still form a feasible subnetwork. Feasibility in this case means that there exists non-zero flux vectors satisfying the steady-state constraint [35, 113]. Flux variability analysis (FVA) is used to identify the removable reaction, with the reaction with the smallest overall flux range selected. This method does not necessarily find the minimum subnetwork with respect to the number of active reactions, and therefore, further reduction could be possible.

Vlassis *et al* proposed the FASTCORE algorithm, which like NetworkReducer, uses linear programming to find subnetworks [135]. However, it does not require FVA and is therefore faster than NetworkReducer. FASTCORE does not find minimum subnetworks and only protected reactions can be specified, and not protected metabolites [113]. Protected reactions and metabolites are those which the user does not want compressed, and therefore effectively ‘lost’ from the network.

Röhl and Bockmayr presented a MILP approach to determine one or more minimum subnetworks. Unlike the other discussed methods, their MILP algorithm ensured minimality of the subnetwork with the active reactions and preserved protected reactions and

metabolites. Their method also enabled several minimum solutions to be found if they existed. The MILP algorithm was also faster than NetworkReducer due to the decreased complexity of the problem [113].

All the discussed methods for network compression reduce large networks to multiple subnetworks via linear programming or MILP methods. However, network compression can be performed by observing the conservation relationships of metabolites alone. This chapter discusses how this simple method can be utilised to reduce the solve time of the MILP approach on Chapter 5, without the loss of any EFMs from the solution.

There are other methods to improve efficiency other than network compression. For example, using sparse matrices will reduce the memory storage required or parallelisation of MILP to allow for multiple cores to be used in the solving of the EFMs. Therefore, this chapter will also discuss how these techniques can be implemented and the effect on solve time.

6.2 Compression Techniques

6.2.1 Sparse Matrices

Reducing a full matrix by the removal of any zeros is the creation of a sparse matrix. The MILP method presented in Chapter 5 used sparse matrices for the equality and inequality constraints within the '*intlinprog*' function. However, this was found to not improve computation. Therefore, sparse matrices were instead used to build up all constraints throughout the MILP setup. This alongside network compression reduces the memory size required.

6.2.2 Integer Cut

The integer cut was needed in the output function and standard MILP setups to prevent repeat solutions clogging up the memory space. However, it has been found that with the addition of equation (5.13), Chapter 5, the integer cut is now redundant. Removal of the integer cut reduces the size of the constraints and aids in the reduction of solve time and memory storage.

6.2.3 Irreversible Reaction Network Compression Methods

Conservation relations of metabolites can be identified as linear dependencies between the rows of the stoichiometry matrix, $\boldsymbol{S}$ [137]. If conservation relations exist then some of the dependent rows of $\boldsymbol{S}$ can be removed such that only independent rows exist [27]. This removal requires the combining reactions to create rows of zeros in $\boldsymbol{S}$. For example, consider a network of 6 reactions and 5 metabolites, Figure 6.1a. This network has the intracellular stoichiometry given in equation (6.1).

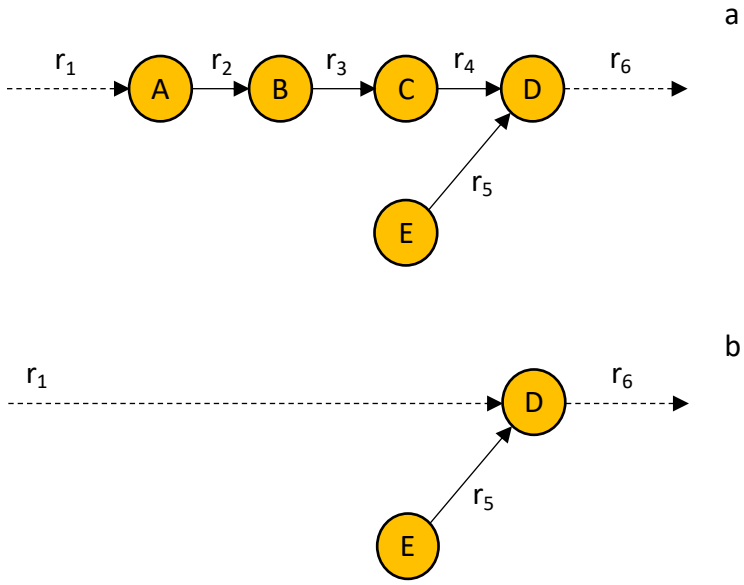

Figure 6.1 a) full network b) reduced network by using conservation relations between metabolites

$$\boldsymbol{S} = \begin{array}{c} \begin{matrix} r_1 & r_2 & r_3 & r_4 & r_5 & r_6 \end{matrix} \\ \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & -1 \\ 0 & 0 & 0 & 0 & -1 & 0 \end{bmatrix} \end{array} \tag{6.1}$$

The first row corresponds to metabolite ‘A’. To make the row a zero-row, columns 1 and 2 can be combined due to ‘A’ being produced by reaction 1 and consumed by reaction 2. This reduces $\boldsymbol{S}$ to,

$$\boldsymbol{S} = \begin{matrix} r_1{+}r_2 & r_3 & r_4 & r_5 & r_6 \end{matrix} \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 & -1 \\ 0 & 0 & 0 & -1 & 0 \end{bmatrix} \tag{6.2}$$

Now metabolite 'A' has an all zero row and is therefore no longer needed to define the network and is therefore removed from the intracellular stoichiometric matrix giving,

$$\boldsymbol{S} = \begin{matrix} r_1{+}r_2 & r_3 & r_4 & r_5 & r_6 \end{matrix} \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 & -1 \\ 0 & 0 & 0 & -1 & 0 \end{bmatrix} \tag{6.3}$$

The next row corresponds to metabolite 'B'. Again, only one reaction reduces this metabolite, and one reaction consumes it. Therefore, columns 1 and 2 can be combined,

$$\boldsymbol{S} = \begin{matrix} r_1{+}r_2{+}r_3 & r_4 & r_5 & r_6 \end{matrix} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 1 & 1 & -1 \\ 0 & 0 & -1 & 0 \end{bmatrix} \tag{6.5}$$

Metabolite 'B' now has a zero row so must be removed,

$$\boldsymbol{S} = \begin{matrix} r_1{+}r_2{+}r_3 & r_4 & r_5 & r_6 \end{matrix} \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & 1 & -1 \\ 0 & 0 & -1 & 0 \end{bmatrix} \tag{6.6}$$

This process is repeated until no more combinations are possible, resulting in,

$$\boldsymbol{S} = \begin{matrix} r_1{+}r_2{+}r_3{+}r_4 & r_5 & r_6 \end{matrix} \begin{bmatrix} 1 & 1 & -1 \\ 0 & -1 & 0 \end{bmatrix} \tag{6.7}$$

The final compressed network is shown in Figure 6.1b. It may be noted that the final compressed network for this example has reduced the number of intracellular metabolites from five to two and the number of considered reactions from six to three. This process is particularly important with cyclic sections of networks as these areas often contain multiple reactions in sequence [138]. For small networks this process can be done manually but larger networks require code.

6.2.4 Irreversible Reaction Network Compression Results

This standard method can be applied to irreversible networks.

6.2.4.1 Hypothetical Cell Network

For example, Figure 6.2a shows a full metabolic network with 7 reactions and 10 metabolites, can be reduced to 3 reactions and 6 metabolites, Figure 6.2b. The new reactions are defined as,

$$R1 = r_1 + 0.5r_3 + 0.5r_7$$

$$R2 = 2r_1 + r_3 + r_5 + 0.5r_6$$

$$R3 = r_1 + r_2 + r_4$$

No solve time or constraint size is quoted as the reduced network for this simple network are the EFMs. This is a unique case due to the simplicity of the network.

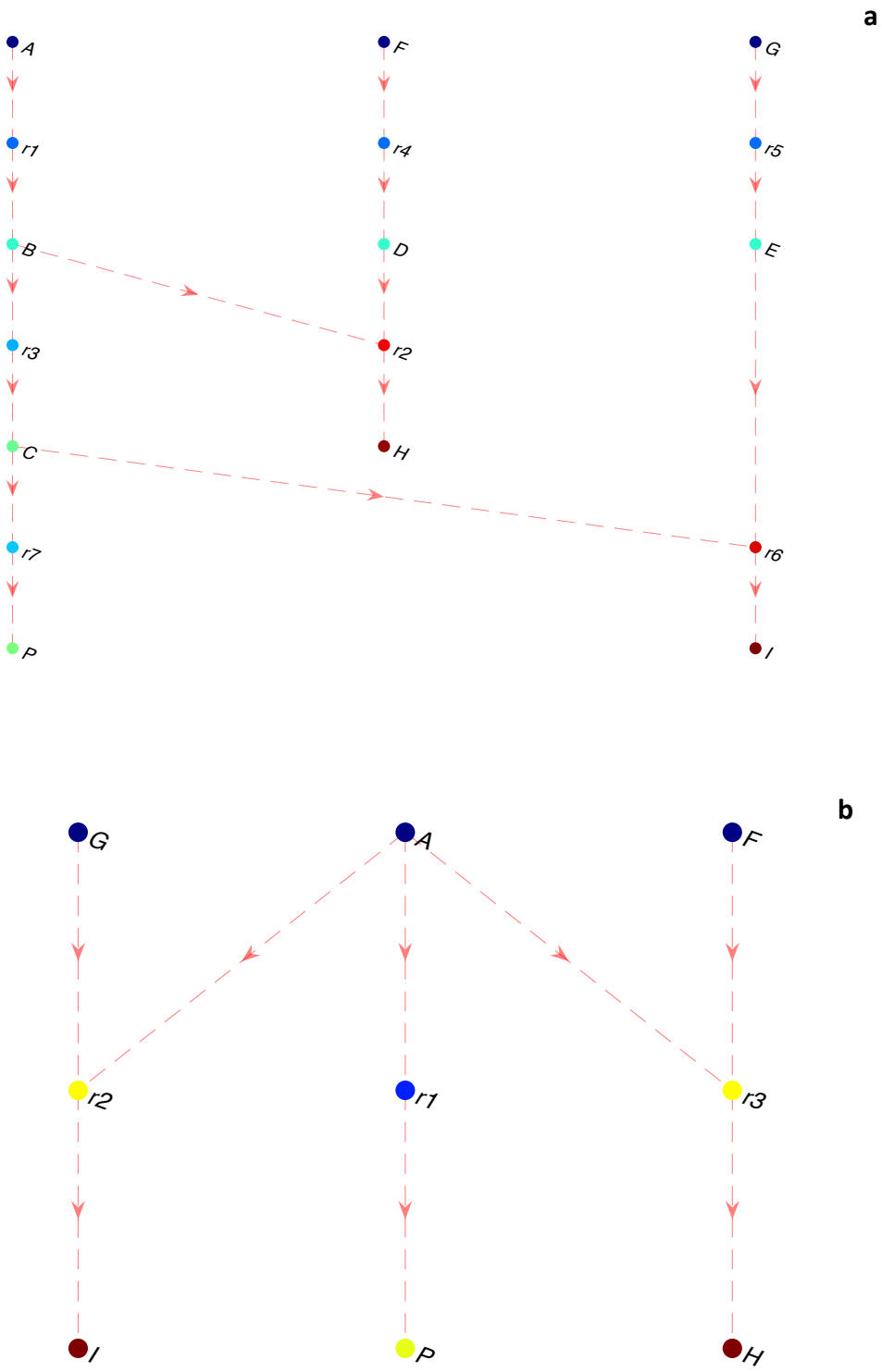

Figure 6.2 a) Simple cell network consisting of 10 metabolites and 7 reactions b) reduced simple cell network consisting of 6 metabolites and 3 reactions.

6.2.4.2 Chinese Hamster Ovary Cell Network

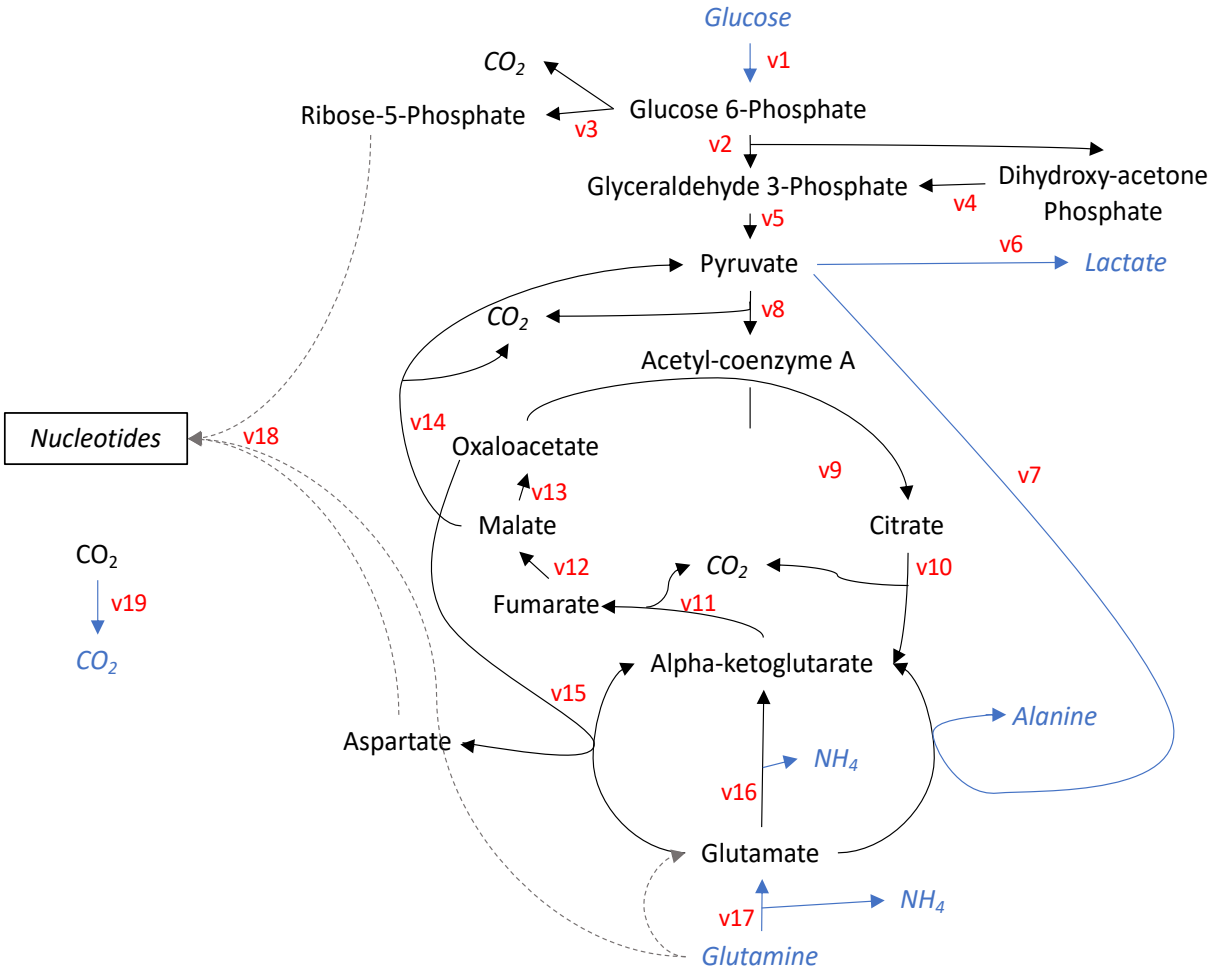

Figure 6.3 Uncompressed simple CHO cell network. Blue indicates the reaction or metabolite is extracellular and black is intracellular

Unlike the previous examples, for most metabolic networks, the compression method merely reduces the size of the network – it does not discover the full set of EFM's. The original Chinese hamster ovary (CHO) cell network consists of 19 reactions and 21 metabolites, Figure 6.3 [90]. However, after network compression, this is reduced to 9 reactions and 11 metabolites, Figure 6.4. The reaction numbers in Figure 6.4 do not correspond to the original reaction numbers due to the adjustment of stoichiometry during network compression,

$$R6 = v_1 + v_2 + v_4 + 2v_5$$

$$R7 = v_{11} + v_{12} + v_{14}$$

$$R8 = v_{11} + v_{12} + v_{13} + v_8 + v_9 + v_{10}$$

$$R9 = v_{11} + 2v_{12} + 2v_{13} + 2v_{15} + v_1 + v_3 + 0.5v_{18}$$

R6 to R9 can be defined as fully coupled; where a non-zero flux for one reaction implies a non-zero flux for the next which is fixed, and vice versa [119]. For example, for reaction 2 in, Figure 6.3 to occur, reaction 1 must have occurred. Therefore, if there is non-zero flux in reaction 1, reaction 2 will also have a non-zero flux fixed by the flux of reaction 1.

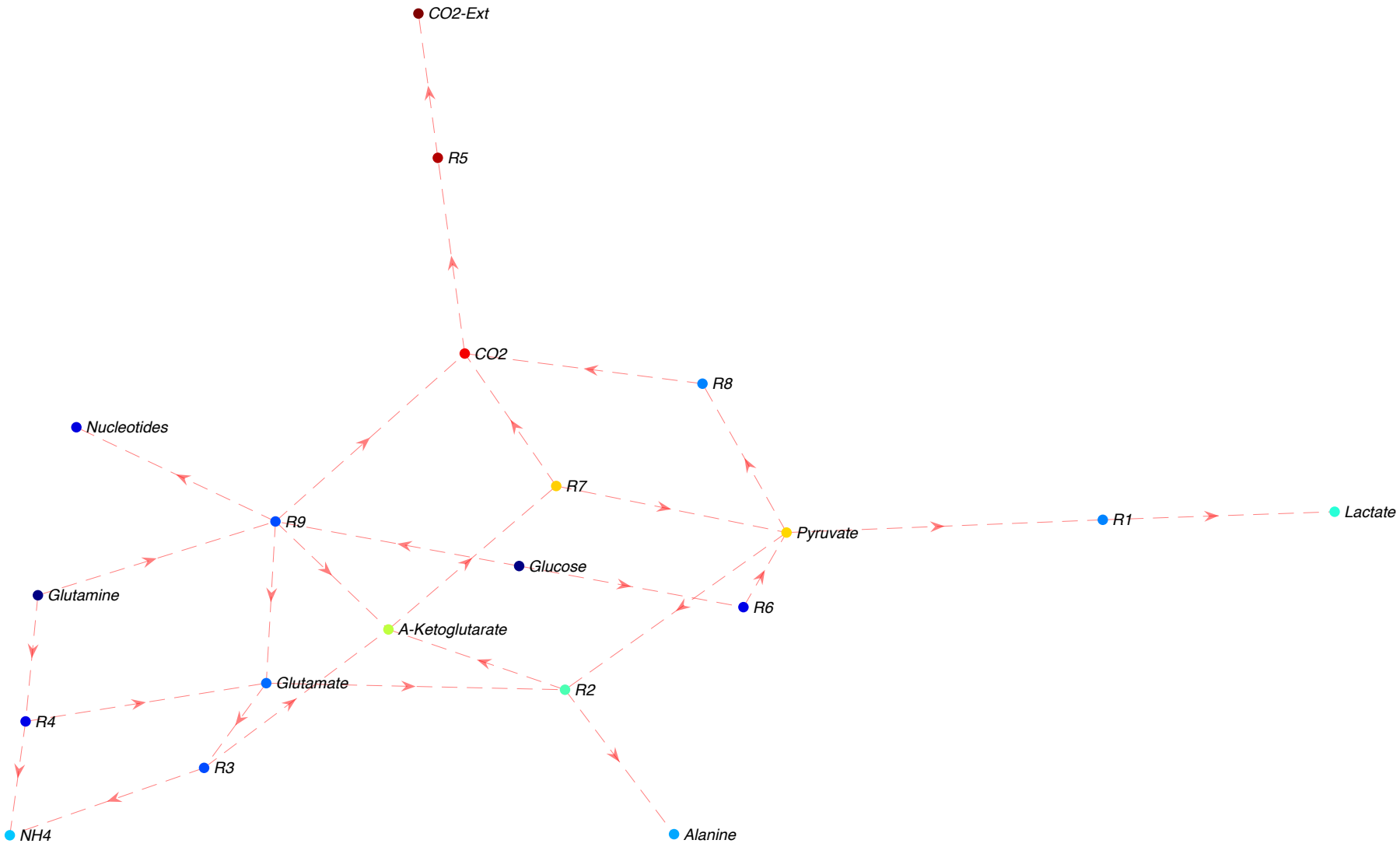

Figure 6.4 Compressed CHO cell network

6.2.4.3 Chinese Hamster Ovary Cell Compression Results

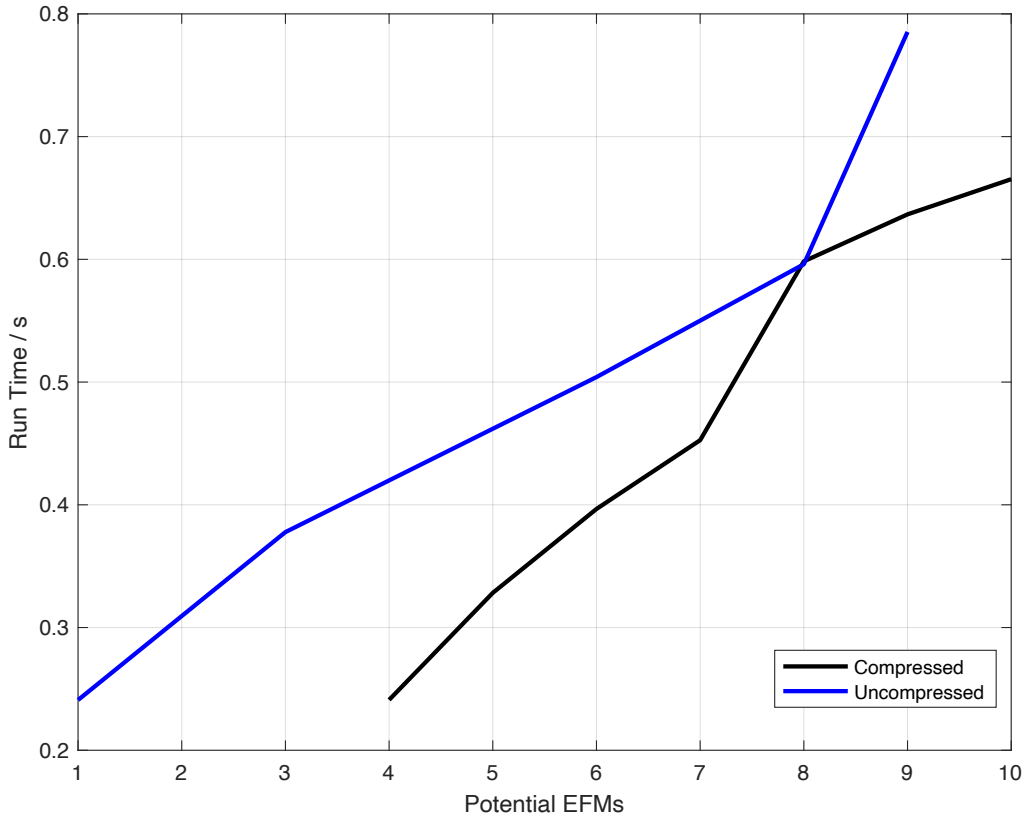

Figure 6.5 Compressed and uncompressed CHO cell run times over EFMs found

Figure 6.5 highlights the reduced solve time achieved via the use of the compressed network. Although this is only a small reduction it highlights that speed up is possible with these adjustments. The compressed network finds 1 extra potential EFM than the uncompressed network in a shorter length of time. This does, however, reduce efficiency with the uncompressed network yielding an efficiency of 1 and the compressed network 0.9. This, however, is a worthwhile reduction as the problem is solved quicker. The uncompressed network has a combined constraint size of 0.0047MB. The compressed network however has a constraint size of 0.0023MB. That is approximately half the storage size required. With larger networks this reduction will become hugely beneficial.

6.2.5 Reversible Reaction Network Compression Methods

The decomposition of reversible reactions expands the problem size for MILP. Therefore, the compression of these reactions is instrumental in reducing the solve time and memory

storage requirements. There are three rules that can be applied to compress these reactions. Firstly, although reversible pairs cannot occur simultaneously, there may be sequential reactions that can if the forward or reverse reaction occurs. For example, Figure 6.6a shows an uncompressed network consisting of 4 reactions and 3 metabolites, with intracellular stoichiometry,

$$\mathbf{S} = \begin{matrix} r_1 & r_2 & r_3 & r_4 \end{matrix} \begin{bmatrix} -1 & 1 & 0 & 0 \\ 1 & -1 & -1 & 1 \\ 0 & 0 & 1 & -1 \end{bmatrix} \tag{6.8}$$

Reactions 1 and 2 cannot occur simultaneously and neither can reactions 3 and 4. However, if reaction 1 does occur in an EFM solution, then reaction 3 must also occur. The same can be said for reaction 4 and 2. Therefore the network can be compressed into 2 reactions and 2 metabolites, Figure 6.6b. These two new reactions will now act as a reversible pair that cannot simultaneously occur in an EFM. The stoichiometry is therefore now,

$$\mathbf{S} = \begin{matrix} r_1 + r_3 & r_2 + r_4 \end{matrix} \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \tag{6.9}$$

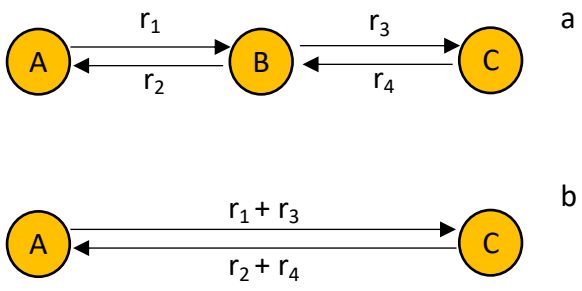

Figure 6.6 a) uncompressed reversible reaction network b) compressed reversible reaction network

Secondly when a metabolite is produced and consumed by a reversible reaction but is only then consumed or produced by one or multiple reactions. For example, in Figure 6.7a metabolite 'B' is consumed by reaction 2 and 3 but is only produced by reaction 1. The stoichiometry is,

$$\mathbf{S} = \begin{matrix} r_1 & r_2 & r_3 \end{matrix} \begin{bmatrix} -1 & 1 & 0 \\ 1 & -1 & -1 \end{bmatrix} \tag{6.10}$$

In any EFM it would not be possible to have reactions 1 and 2 as they are a reversible pair. However, if reaction 1 occurs then reaction 3 must also. Therefore, the network can be compressed to that shown in Figure 6.7b. The compressed network still contains a reversible reaction pair, but the number of reactions is reduced from 3 to 2. Reactions 1 and 3 combined effectively skip out metabolite 'B', giving the new stoichiometry,

$$\mathbf{S} = \begin{matrix} r_1 + r_3 & r_2 \\ \end{matrix} \begin{bmatrix} -1 & 1 \\ 0 & -1 \end{bmatrix} \tag{6.11}$$

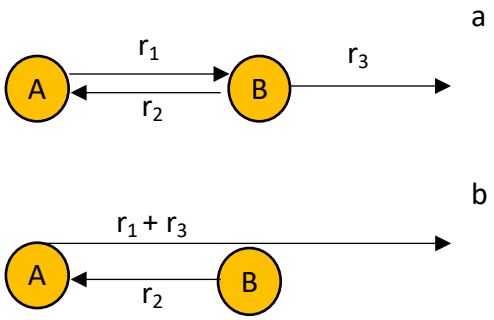

Figure 6.7 a) uncompressed network with one reversible reaction pair b) compressed version of the network

The final rule that can be applied regards if a metabolite in a chain is produced and consumed by a reversible reaction, Figure 6.8a [138]. In this network 'A' acts as the substrate and reaction 5 continues in the network to eventually produced an extracellular product. There exists only one EFM for this network $\boldsymbol{E_1} = [1\ 1\ 0\ 0\ 1]^T$, confirmed via efmtool and the MILP method presented. The reversible reaction comprising of reactions 3 and 4 is effectively redundant. Therefore, the network can be compressed to that in Figure 6.8b. This has reduced

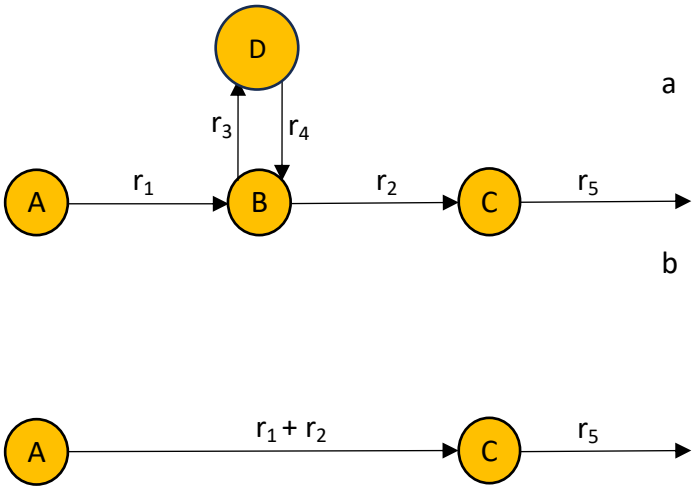

Figure 6.8 a) uncompressed network containing redundant reversible reaction b) compressed network with redundant reversible reaction removed

the number of reactions from 5 to 2 and the number of metabolites from 4 to 2. Applying this technique to larger networks will reduce the search space for MILP.

6.2.6 Reversible Reaction Network Compression Results

6.2.6.1 Yeast Core Network

A simple yeast core network is given in Figure 6.9 [122]. The network consists of 40 reactions, 18 of which are reversible pairs, and 26 metabolites. Compression without considering reversible reactions reduces the network to 34 reactions and 20 metabolites. Compressing the reversible reactions reduces the reactions to 29 and metabolites to 18.

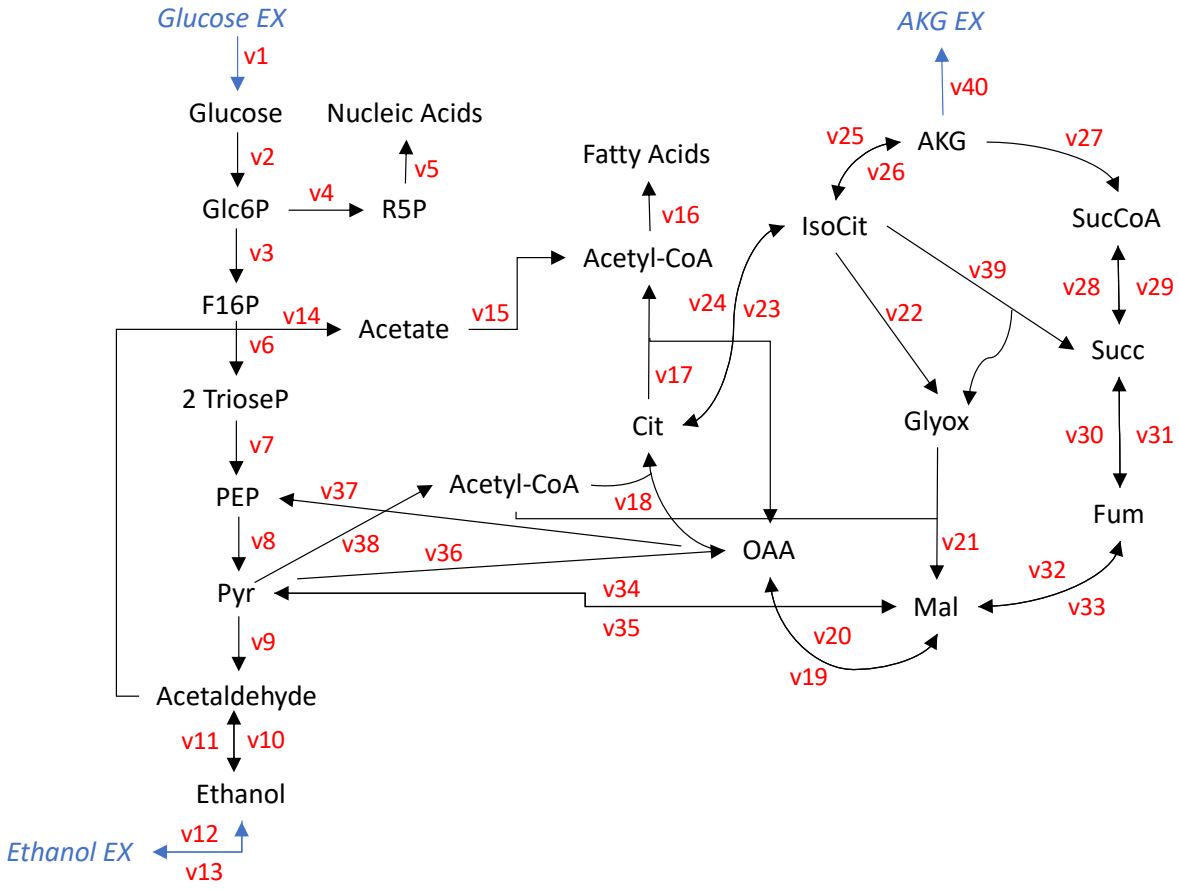

Figure 6.9 Simple yeast core network. Reaction numbers in red and extracellular metabolites in blue

The compressed network is given in Figure 6.10 with a closer figure exploring the cyclic reactions given in Figure 6.11. The new reactions which are fully coupled for this network are as follows,

$$R3 = v_{10} + v_{12}$$

$$R4 = v_{11} + v_{13}$$

$$R16 = v_{27} + v_{28}$$

$$R17 = v_{29}$$

$$R18 = v_{30} + v_{32}$$

$$R19 = v_{31} + v_{33}$$

$$R27 = v_1 + v_2 + v_3 + v_6 + v_7$$

$$R28 = v_1 + v_2 + v_4 + v_5$$

$$R29 = v_{14} + v_{15}$$

The most important coupled reaction is $R27$. This reduces 5 reactions to 1, thus reducing the search space for MILP. Glucose is one of two main nutrients for this network, the other being ethanol. Therefore, EFMs will have to be generated from the reaction that consumes it, reaction 1 in Figure 6.9. By compressing the reactions that follow, you are guaranteed to reduce solve time. Cyclic reactions lead to multiple routes for MILP to follow. Although the reductions are small by coupling the reactions in these cycles it helps with the search time, for example $R18$ and $R19$.

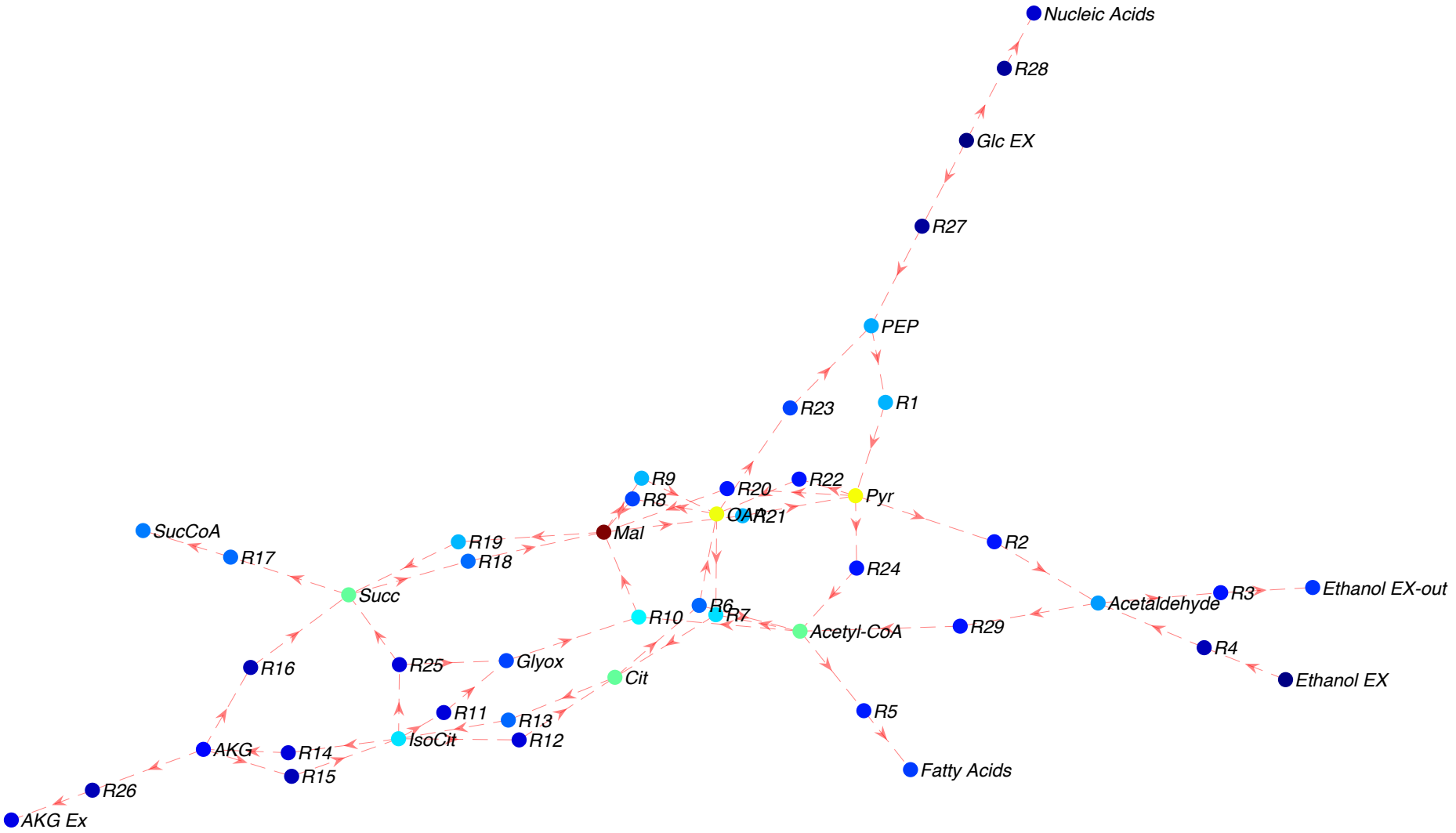

Figure 6.10 Compressed yeast core network

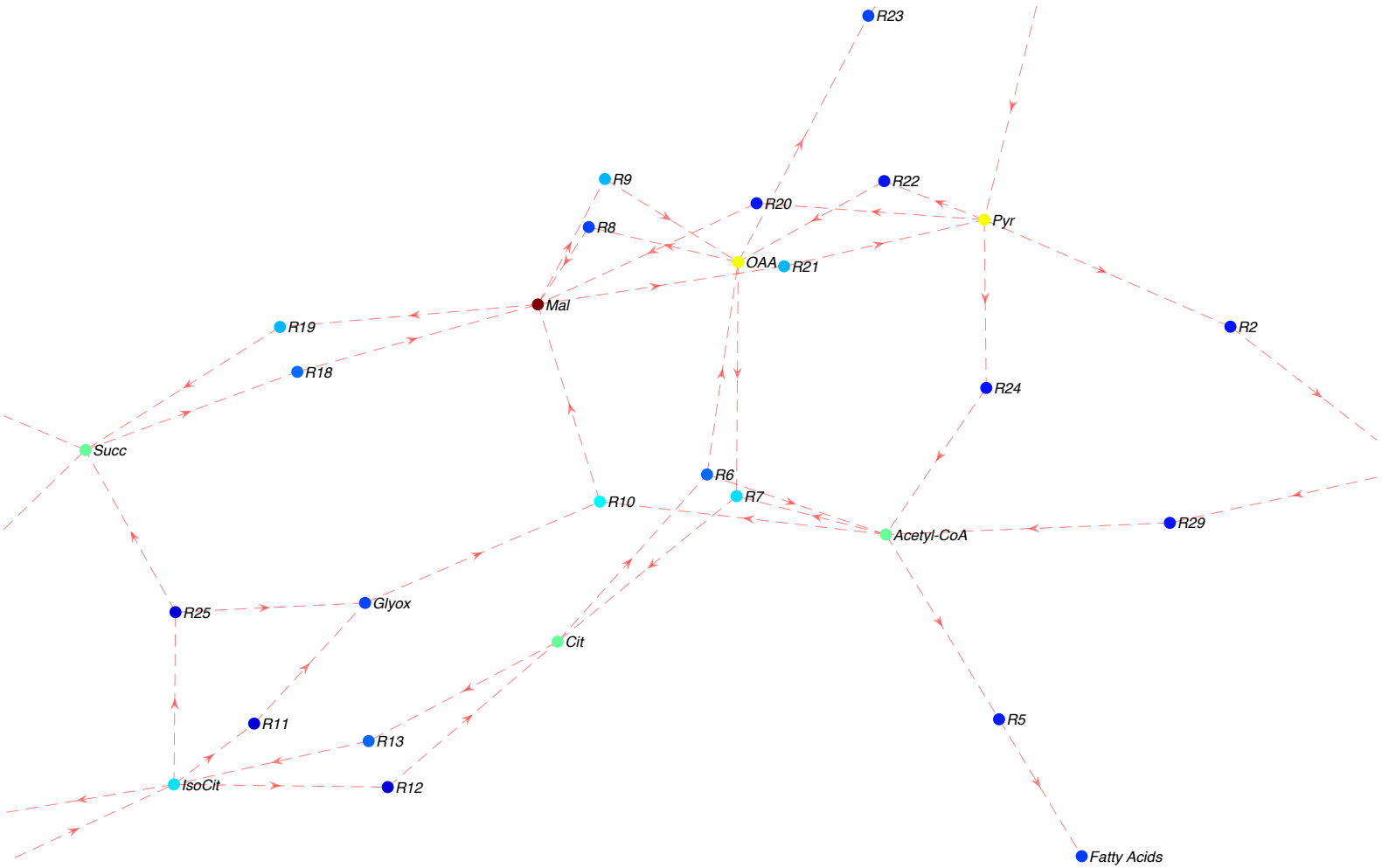

Figure 6.11 Zoom in of compressed yeast core network

6.2.6.2 Yeast Core Compression Results

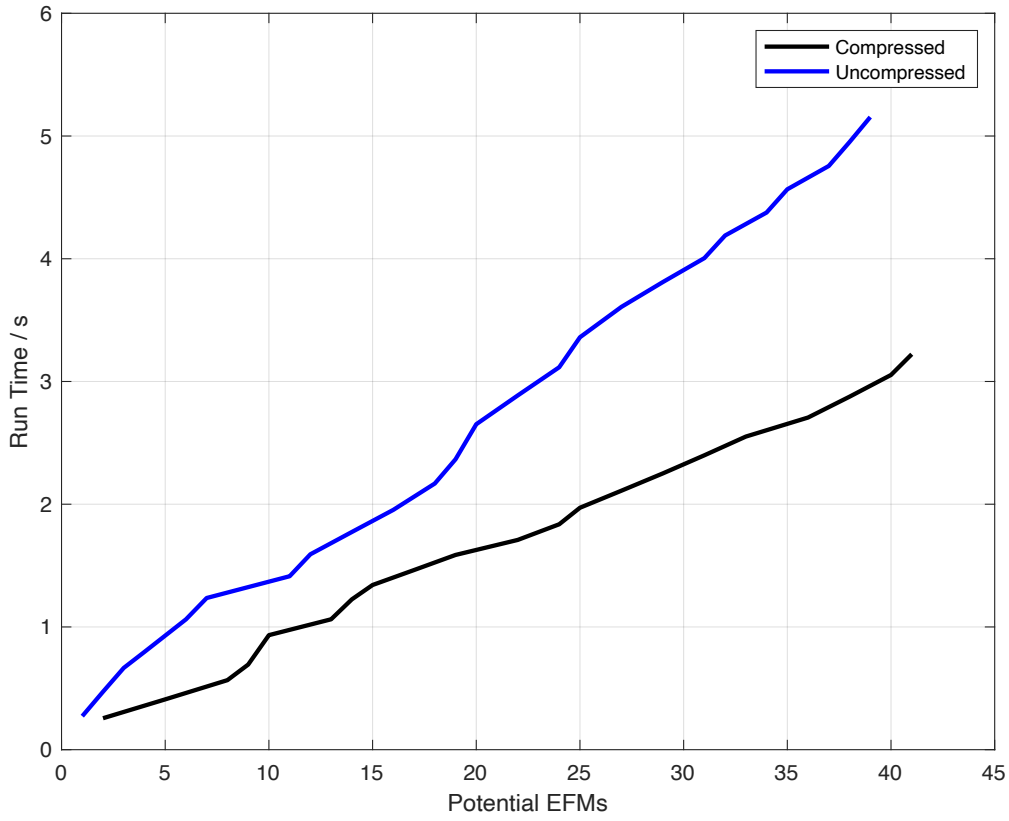

Figure 6.12 Compressed and uncompressed yeast core cell run times over EFMs found

The improvements made by compressed the yeast core network are shown in Figure 6.12, highlighting that more potential EFMs are found, again in a shorter time. The run time for the compressed network is 1.97s, approximately a 50% reduction in solve time compared to the uncompressed network. The constraint memory sizes have also decreased, the equality constraint decreased from 0.001928MB to 0.001432MB and the inequality constraint from 0.02004MB to 0.0142MB.

6.3 Compressing the *E. coli* Core

Utilising all the discussed compression methods the *E. coli* core was reduced from 114 reactions, of which 39 are in reversible pairs, and 92 metabolites to 90 reactions, of which 30 are reversible pairs, and 84 metabolites. This drastically improved the number of EFMs achievable in 10 minutes and 1 hour, Table 6.1. In 10 minutes compared to the previous MILP setup (results in Chapter 5 section 5.5.1) there was a 436% increase in the number of EFMs

found. The increase for 1 hour was 302%. Iterations for both these runs had to be increased to 250, as 100 iterations was solved prior to the allocated time. Efficiency for both these runs also improved on the results reported in Chapter 5 on average by 12%. This demonstrates that network compression, removal of integer cuts and sparsity of matrices reduces solve time for large networks. Therefore, compression is required for MILP in enumerating EFMs at large scale and creates the opportunity for genome scale in the future.

Table 6.1 E. coli network efficiency and number of EFMs found in 10 minutes and 1 hour

Run Time	**Number of EFMs found**	**Efficiency**
10 min	817	0.8988
1 hour	1018	0.8985

As an example of the improved run times both the uncompressed and compressed network were run for a maximum of 50 iterations, Figure 6.13. The compressed network finds slightly more potential EFMs than the uncompressed network, however, the reduced solve time for the compressed network is the key takeaway.

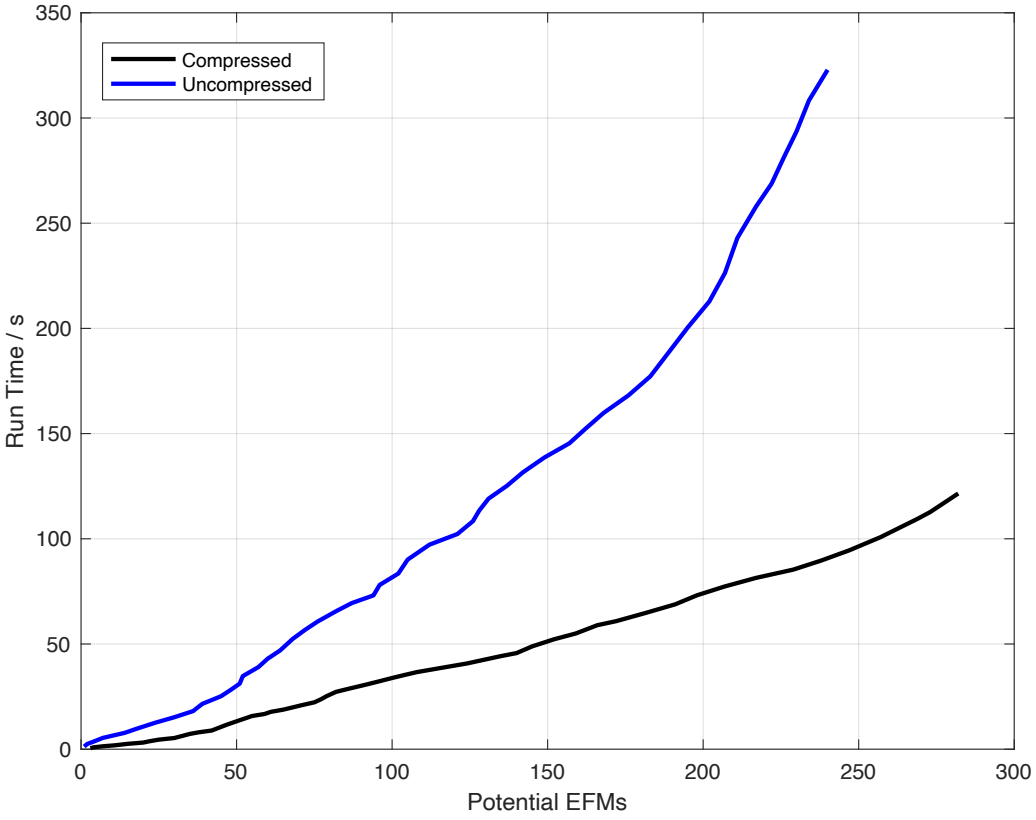

Figure 6.13 E. coli core compressed and uncompressed network run times over EFMs found in 50 iterations

6.4 Parallelisation

Solving larger networks using efmtool requires the use of multiple cores [30]. Using multiple computers or multiple cores within one computer will also improve the efficiency of MILP in enumerating EFMs. However, efmtool is not the only tool to report on the use of parallelisation methods. Klamt *et al* presented a case to split the computation into two independent processes [27]. Firstly, a reaction (n) is chosen and (i) the set of EFMs which involve reaction n are computed, (ii) then all EFMs not containing reaction n are computed. The complete set of EFMs consists of the solutions obtained in (i) and (ii). This computation process is possible due to steps (i) and (ii) being independent [27].

Klamt *et al's* work also discussed the splitting of EFM enumeration into 2^b independent processes, where the process above is applied recursively leading to a binary tree with b layers and 2b leaves, Figure 6.14. The root node is the complete problem to find all EFMs,

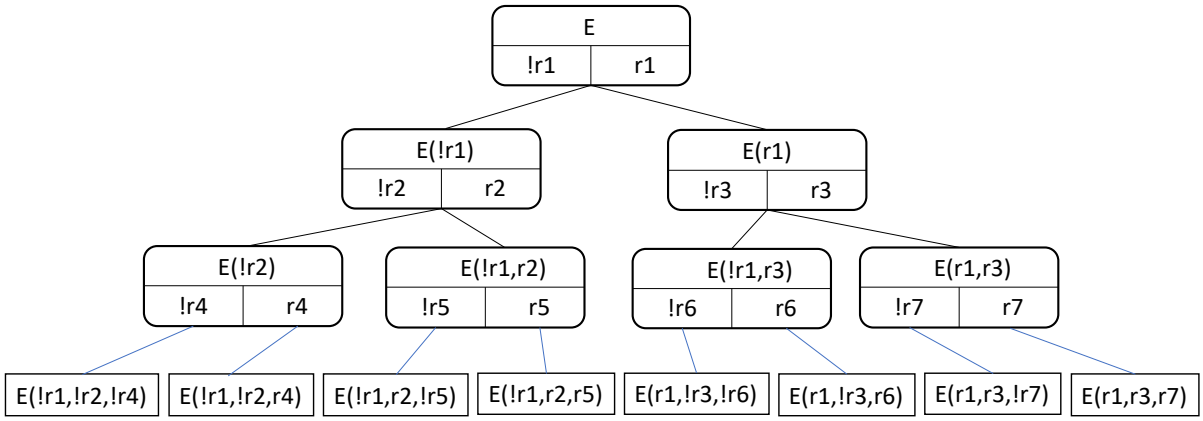

Figure 6.14 Splitting the computation of EMs into independent sub-tasks

intermediate nodes are the subtasks that be computed independently, and each leaf is a final subtask. Each intermediate node has a different reaction, n, which yields EFMs that contain and do not contain it. Collecting all 2^b results together will provide the full EFM set.

Subtasks can be performed on multiple computers or cores. In the MILP reactions can be turned on or off through use of their associated binary variable were specifying that a particular term is either 1 or 0 indicating if they should or should not exist within the EFM set.

6.4.1 Yeast Core Network

For example, for the yeast core network reactions 14 and 18 can be turned on and off in multiple combinations, Table 6.2. Solving these four combinations separately reduces the solve time and still finds all 39 EFMs. Reactions 14 and 18 are not a unique case, the same method was applied to reactions 9 and 10, Table 6.3. Again all 39 EFMs were found with each combination running in an average of 1.12s.

Table 6.2 Reactions 14 and 18 combinations and the EFMs found

Reaction combination	Solve Time / s	EFMs
R14 on R18 on	1.2192	10
R14 on, R18 off	0.6874	6
R14 off, R18 on	1.6521	12
R14 off R18 off	0.9189	11

Table 6.3 Reactions 9 and 10 combinations and the EFMs found

Reaction combination	Solve Time / s	EFMs
R9 on R10 on	1.4119	14
R9 on, R10 off	1.1518	7
R9 off, R10 on	1.1061	8
R9 off R10 off	1.0030	10

6.4.2 E. coli Core Network

A similar approach has been taken with the *E. coli* core. Four reactions have been chosen due to their locations in the network either connecting central section's e.g., connecting the TCA cycle to the network, or due to the fact they act as intermediates in long chains of reactions. These reactions were then forced on and off in the MILP code to find how many EFMs could be detected in within a maximum of 50 iterations. The results for the various combinations of reactions 15, 27, 44 and 74 are given in Table 6.4. These multiple combinations yielded 1077 EFMs, a greater value than that obtained by running the code for an hour. This is due to the reduced search space created by on/off reaction constraints. Some reaction combinations yielded no EFMs, for example R15 and R27 on and R44 and R74 off. Rather than holding up the entire computational power searching for the possibility of EFMs in conditions were there are none, parallelisation will allow for non EFMs to be found and move on whilst searching for other EFMs on different cores or computers.

Further reaction combinations could be applied to obtain more EFMs. All the EFMs found by these individual searches would require post-processing to ensure that each EFM is non-decomposable into another. However, this again proves the benefits of parallelisation of MILP to obtain more EFMs. Any reaction combinations could be run on multiple computers or cores, reducing the search time for EFMs with MILP at large scale. In future, this would allow for genome scale to be solved.

Table 6.4 Reactions 15,27,44 and 74 combinations and the EFMs found. Red indicates off and green indicates on

Reaction combination	Solve Time / s	EFMs
R15, R27, R44, R74	124.3353	246

R15, R27, R44,R74	105.1657	166
R15, R27, R44,R74	0.1931	0
R15, R27, R44,R74	0.1859	0
R15, R27, R44,R74	0.1743	0
R15, R27, R44, R74	0.1606	0
R15, R27, R44, R74	0.1740	0
R15, R27, R44, R74	0.1538	0
R15, R27, R44, R74	105.6646	140
R15, R27, R44, R74	97.5701	93
R15, R27, R44, R74	87.8148	70
R15, R27, R44, R74	80.5641	218
R15, R27, R44, R74	0.1780	0
R15, R27, R44, R74	64.7635	112
R15, R27, R44, R74	86.4769	32
R15, R27, R44, R74	0.1782	0
Total number of EFMs found		1077

6.5 Application of Flux Data

Flux data can be utilised when finding EFMs to reduce the search space. Flux indicates what reactions are in operation, with a 0-flux indicating the reaction is not being used. The application of this data when finding EFMs will therefore reduce the branch and bounding steps required by the MILP. For example, for the network in Figure 6.10 if it was found that the flux through R2 was equal to 0, this reaction could be removed from the search space. This reduces the EFMs from 39 to 25 and the efficiency is 1. These EFMs are found in 1.92s. Therefore, direct use of flux data to set 0-flux reactions to off in the MILP method will reduce the search space and reduce solve time.

In Chapter 3, section 3.3.1 an underdetermined CHO cell was used to find the range of fluxes via FVA to maximise lactate production. The minimum FVA results, Figure 3.8, showed only 5 reactions were required, however setting these reactions to be on and all other reactions off generated no EFMs in 0.89s. This shows a case where application of flux data over constrains

the search space for EFMs. Therefore, flux data can be applied to help with the search of EFMs, but it must be done with realistic cases only. The minimum flux data generated by FVA is a hypothetical case which is unlikely to occur. Applying a variation of FVA results from the minimum results to the maximum is a better method. For example, for the CHO cell, if one extra reaction is added to the 5 minimum occurring reactions, 1 EFM is found in 0.25s.

6.6 Conclusion

The MILP approach of branch and bound causes memory storage clogging and reduced solve time for EFM enumeration, particularly at a large scale. However, this chapter has shown that improvements can be made to alleviate this issue. Firstly, the removal of unnecessary integer cuts and introducing sparse matrices throughout the code reduced the strain on the memory. Although memory size is small for the networks discussed, genome scale requirements would lead to large data sets that have to be stored throughout computation. Therefore, reducing the number of constraints and how the matrices are saved is crucial to the future of MILP in EFM enumeration. Secondly, compressing networks in some cases halved the solve time. Compression can be performed on reversible or irreversible reactions, reducing the number of metabolites and reactions in a network's stoichiometry. Combining these techniques made it possible to solve over a 1000 EFMs in the *E. coli* core, a 302% increase on the method presented in Chapter 4. Reducing the search space was also found to be possible by applying flux data. Zero fluxes indicate that reactions are not in operation, and this can be used to prevent EFMs containing these reactions being found via MILP. However, flux data needs to be carefully used with the understanding that FBA and FVA are ideal, non-realistic cases. Their results offer a starting point in reducing the search space. Finally, this chapter discussed the benefits of running MILP code in parallel to one another to speed up EFM enumeration. This technique would future-proof MILP so it could be used on much larger networks, and potentially genome scale, to find EFMs. Overall, this chapter has emphasised the improvements needed to enable MILP to find more EFMs in a shorter period, whilst highlighting future methods that could be applied to apply the technique at genome scale.

Chapter 7 The *E. coli* Cell

7.1 Introduction

The production of recombinant proteins in microbial systems changed the vaccine industry drastically. There is no longer a need for vast amounts of animal and plant tissues or large volumes of biological fluids to produce the desired quantities of the proteins [22]. Large-scale trials have shown that <50% of bacterial proteins and <15% of non-bacterial proteins can be expressed in Escherichia coli (*E. coli*) [21]. This demonstrates the versality of the cell and emphasises the importance of ensuring high yields.

This chapter presents the data collected by GlaxoSmithKline (GSK) at their Rixenstrat, Belgium site. The key groups within the core metabolism will be discussed highlighting areas of interest. Flux analysis will be presented, the generation of elementary flux modes (EFMs) and the limitations of the data highlighted.

7.2 Key Groups of the Core Metabolism

The *E. coli* core (reduced set of reactions representing the key areas of the network) metabolism is often used as a start point before moving onto the genome. Within the core metabolic network there exists 11 groupings of biochemical reactions [41]. These groupings help split the 95-reaction model into easier to understand parts, but it should be noted that some reactions are used in multiple groupings. Each grouping also plays a key role in the cellular function and therefore can be applied to many other cell's metabolic networks.

7.2.1 Glycolysis

The glycolysis set consists of 10 reactions: converting sugars into precursors for biomass. The reaction set terminates with pyruvate. Some small amounts of ATP (adenosine triphosphate) and NADH (Nicotinamide adenine dinucleotide) are also formed, which act as key global metabolites within the network. Most reactions rely on the supply of ATP, and in cases where this supply is low, biosynthesis suffers [139]. The pyruvate produced in this grouping is a key component in many other reactions across the network.

7.2.2 Pentose Phosphate Pathway

The pentose phosphate pathway (PPP) group consists of 8 reactions. Two biosynthetic precursors are created in this grouping: α-D-ribose 5-Phosphate and D-erythrose 4-phostphate. These precursors can be made by one of 2 parallel routes, one oxidative and the other nonoxidative [41]. If the conditions are anaerobic there is greater flux through the non-oxidative route [41, 140].

7.2.3 Tricarboxylic acid cycle

The tricarboxylic acid (TCA) cycle's function changes based upon the environment. During aerobic growth on 6-carbon sugars, the TCA cycle's function is to produce the precursors oxaloacetate, 2-oxoglutarate (-ketoglutarate) and succinyl-CoA. However, during anaerobic growth the TCA cycle functions as two separate pathways. In general, it acts as the final route for the oxidation of fuel molecules, such as amino acids and carbohydrates [141].

7.2.4 Glyoxylate Cycle, Gluconeogenesis and Anaplerotic Reactions

The glyoxylate cycle is used instead of the TCA cycle to bypass reactions that lose carbon in the form of carbon dioxide. Therefore, the glyoxylate cycle does consist of some of the same reactions as the TCA cycle. The reversal of the glyoxylate cycle is known as gluconeogenesis [38].

To replenish the intermediates of the TCA cycle, which are used for biosynthesis, anaplerotic reactions are required. The TCA cycle can oxidise acetate to carbon dioxide without any consumption or production of intermediates. The intermediates (e.g., oxaloacetate) are consumed in the production of macromolecules (large molecule such as a protein) [41].

7.2.5 Electron Transport Chain, Oxidative Phosphorylation, and Transfer of Reducing Equivalents

The electron transport chain and oxidative phosphorylation are used to produce most of the ATP for the cell under aerobic conditions. The electron transport chain moves protons (H+) from the cytoplasm via the cytoplasmic membrane (plasma membrane) into the periplasmic space [142]. The cytoplasmic membrane is impermeable to protons and electrons (OH-) so a

difference in electrical charge will occur across it. This is the thermodynamic potential difference [41].

The potential difference drives the reactions with protons assumed to move to the extracellular medium. This assumption is since the pH of periplasm and extracellular medium is near the same [41, 143]. The reactions involved in these groupings are given in equations (7.1) to (7.8) [41]. Equation (7.5) is an entirely hypothetical reaction that is only required due to the core network not encompassing all reactions that consume ATP.

NADH dehydrogenase (uniquinone-8 and 3 protons)	$4h + nadh + q8 \rightarrow 3h[e] + nad + q8h2$	(7.1)
Cytochrome oxidase bd (uniquinol-8: 2 protons)	$2h + \frac{1}{2}o2 + q8h2 \rightarrow h2o + 2h[e] + q8$	(7.2)
Oxygen transport via diffusion	$o2[e] \leftrightarrow o2$	(7.3)
ATP synthase (4 protons: ATP)	$adp + 4h[e] + pi \leftrightarrow atp + h2o + 3h$	(7.4)
ATP maintenance requirement	$atp + h2o \rightarrow adp + h[e] + pi$	(7.5)
Adenylate kinase	$amp + atp \leftrightarrow 2adp$	(7.6)
NAD(P) transhydrogenase	$2h[e] + nadh + nadp \rightarrow 2h + nad + nadph$	(7.7)
NAD transhydrogenase	$nad + nadph \rightarrow nadh + nadp$	(7.8)

7.2.6 Fermentation

The fermentation process produces end products from sugars within the cell [38]. Substantial amounts of carbon dioxide and hydrogen are produced in this set of reactions. The flux can vary throughout the network based upon the pH of cell at the time. All end products of fermentation do leave the cell, via a concentration gradient, thus transferring a proton from the cytoplasm to the periplasmic space.

7.2.7 Nitrogen Metabolism

Nitrogen is the 4th most abundant metabolite in the *E. coli* metabolism. Nitrogen enters the cell as either ammonium (nh4[c]), or as a moiety within glutamine (glu-L[c]) or glutamate (gln-L[c]) [38].

7.3 The Process Data

Six sets of data were supplied by GSK. The process data can be broken down into 5 parts: sampling, feed, off-gas, product composition and fermentation. The average fermentation volume at time 0hrs (after inoculation) was 9023.8mL with fed-batch being initialised at ~21hrs. The airflow was 20NLmin^{-1}, pH was 7 and the temperature at 0hrs was 37^{o}C, decreasing to 28^{o}C at 66hrs. A total of 19 samples were taken during the experiments, with 10 being taken between 40 and 48 hours. This time range highlights the transitional phase of the cell's life.

An antifoam (3mL of SAG47) and antibiotic (4.5mL kanamycin) were bolus fed to the cell cultures at 0hrs. A base feed of ammonia (density: 0.88gmL^{-1}) was also added to the culture every 15 minutes. To ensure the cells could successfully grow, glucose (density: 1.20106 gmL^{-1}) was required. This was added every 15 minutes and consisted of: beta-D-glucose (2775.31mmolL^{-1}), dipotassium phosphate (122.85 mmolL^{-1}), monopotassium phosphate (76.72 mmolL^{-1}) and L-Isoleucine (15.25 mmolL^{-1}).

Figure 7.1 and Figure 7.2 provide off-gas data for oxygen uptake rate (OUR) and carbon dioxide excretion rate (CER). All growth cultures have an OUR and CER excretion rate peak at ~20hours, except fermentation MME17 due to measurement errors. In the first 20 hours the cells are grown in batch. They, therefore, consume available sugar as fast as possible leading to an exponential growth. Once the sugar has been exhausted, the OUR and CER drop drastically. The decrease in OUR leads to an increase in the dissolved oxygen, signalling that the fed-batch must be initialised. The difference in the OUR and CER in MME17 can be put down to measurement error initially, which was highlighted in the data set provided for this experimental run.

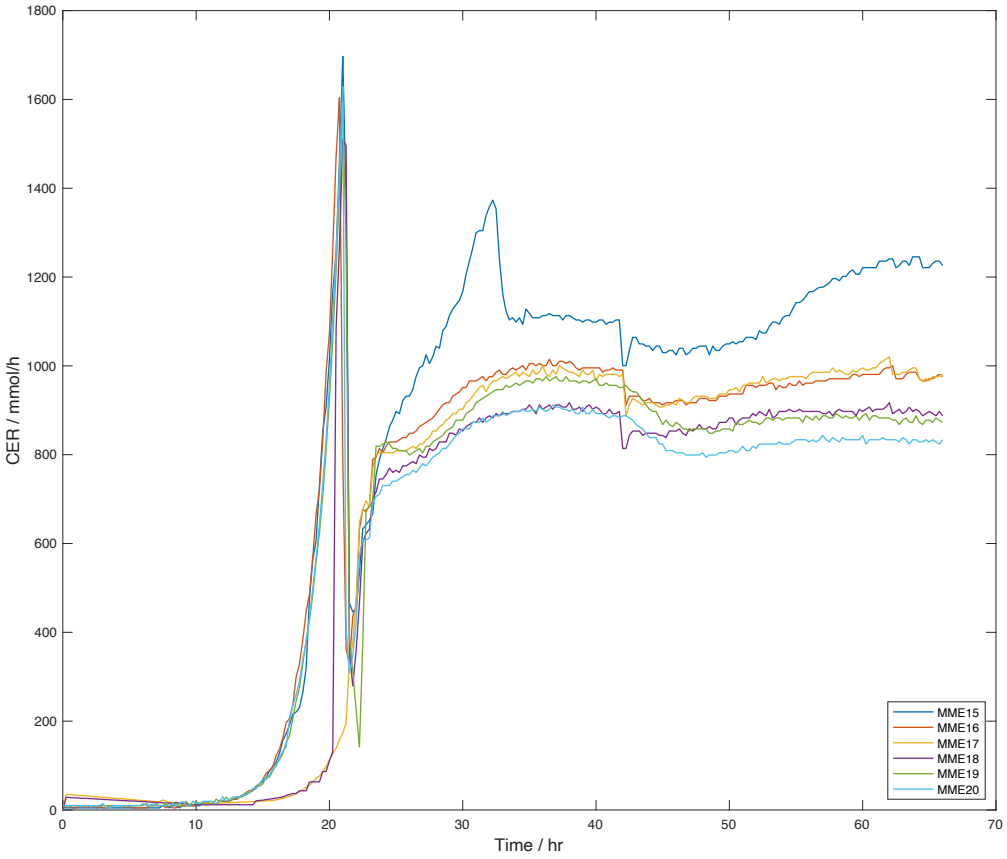

Figure 7.1 Carbon dioxide excretion rate for all cell cultures

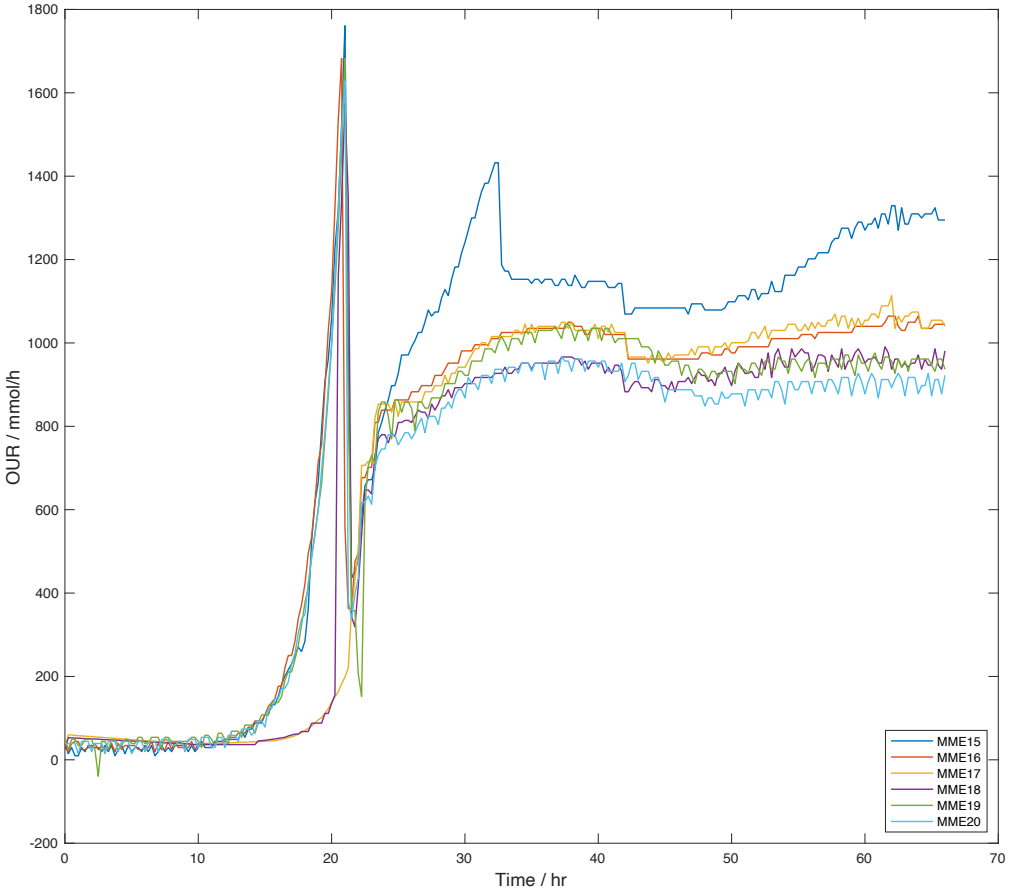

Figure 7.2 Oxygen uptake rate for all cell cultures

The product compositions of each of the cell cultures are given in Figure 7.3. Environmental and operating conditions were maintained as close as possible for all cultures; therefore, the differences are not due to these factors. The variation in the compositions is due to the antigens present within the cells. All fermentations begin with an *E. coli* strain with genome only and no plasmid. A plasmid is a small circular DNA molecule found in bacteria and some other microscopic organisms that can that can replicate autonomously [144]. One of three different plasmids, containing the coding sequence for the respective antigen, is added to a different 'empty' *E. coli* strain. Integration of the plasmid to the strain is promoted with a treatment such as temperature change or electric shock. A single clone of the *E. coli* cell is selected where integration was successful to create the inoculum. Plasmid integration is known as bacterial transformation [145].

An antigen induces an immune response within the body. Production of these antigens is vital for producing vaccines and in the combat of cancer. The process data consist of three antigens: WT1 (Figure 7.3a), M72 (Figure 7.3b) and F4co (Figure 7.3c).

WT1, also known as Wilms' tumour gene 1, has been found to be a useful target antigen in tumour specific immunotherapy in the human invitro system [146]. Oka *et al* proposed that the creation of a WT1-based T cell therapy and vaccine would help against a variety of malignant cancers [147].

M72 has been used in a vaccine produced by GSK against *Mycobacterium tuberculosis.* The vaccine was found to elicit an immune response and gave participants to the trial at least three years protection against progression to pulmonary tuberculosis [148]. F4co has also been used in a vaccine, but in the reduction of viral load in those infected with HIV [149].

Although this process data only encapsulates three antigens, the importance of these antigens is clear. More importantly it shows the versatility of *E. coli* and why there is such interest in optimising its growth in suspension culture for pharmaceutical/therapeutic applications.

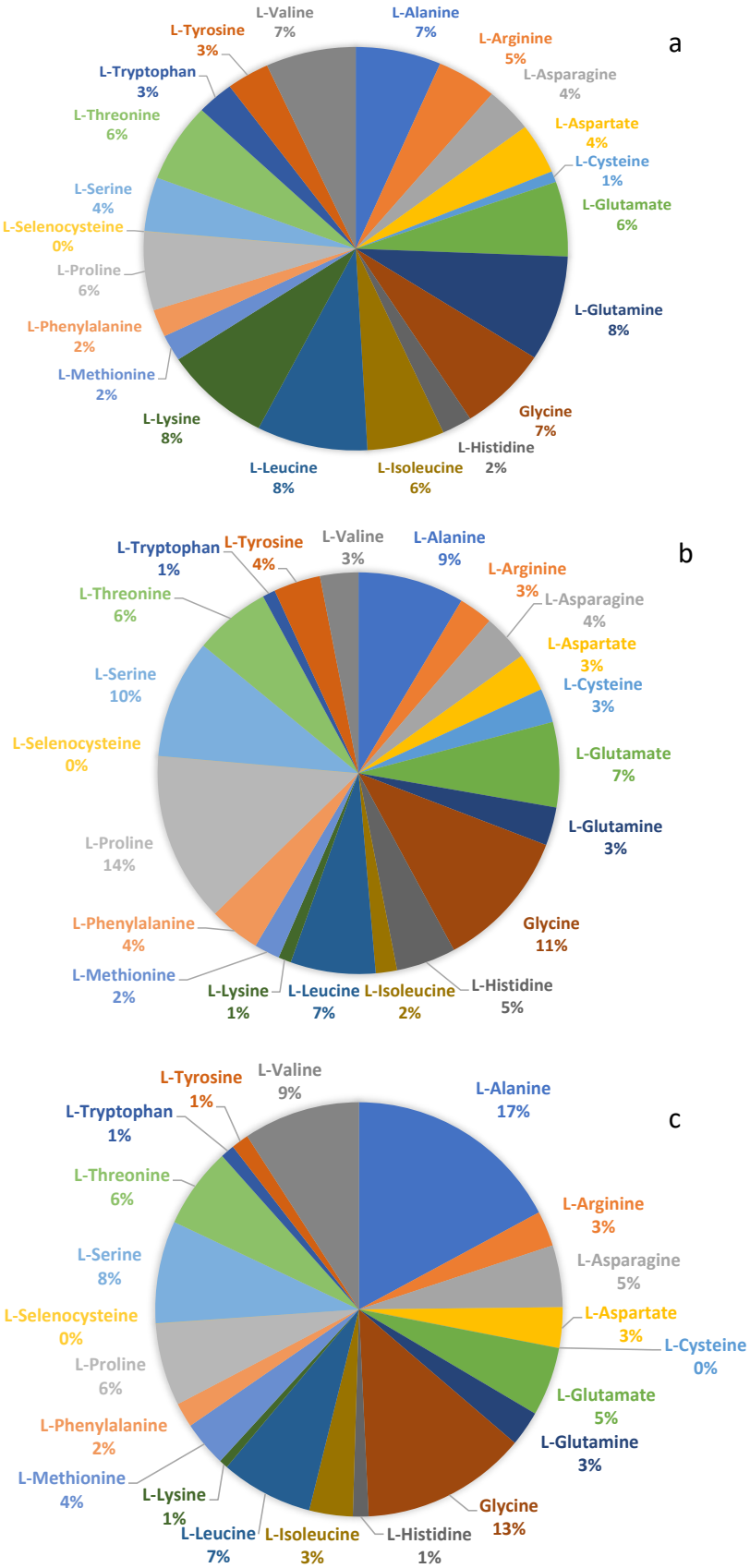

Figure 7.3 Product composition comprising of the amino acid molar composition for a) WT1 grown in MME15 and MME16 b) M72 grown in MME17 and MME18 and c) F4co grown in MME19 and MME20

7.3.1 Biomass Composition

Biomass consists of 20 amino acids, RNA and DNA. Figure 7.4 shows the biomass concentration per fermentation across the sample time. Biomass data across the 6 fermentations all show similar profiles, with fermentations with the same antigen correlating most.

In addition to the similarity in biomass composition, the more in-depth amino acid comparison shows the consistency across all fermentations, Figure 7.5. Details of RNA and DNA composition are given in Figure 7.6.

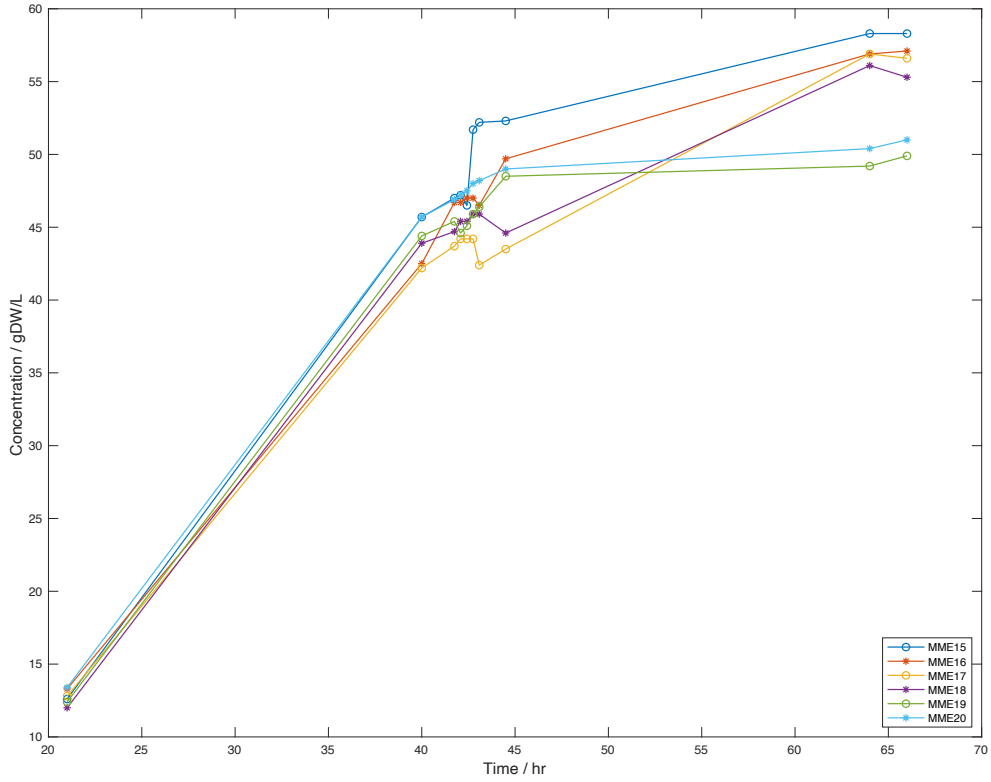

Figure 7.4 Biomass concentration across the fermentation time with sampling points shown

7.3.2 Absolute Quantification

48 metabolites were measured throughout the sample time. These were either classed as extracellular, intracellular, or low-intracellular. Low intracellular showed that the metabolite amount in the sample was less than 5%.

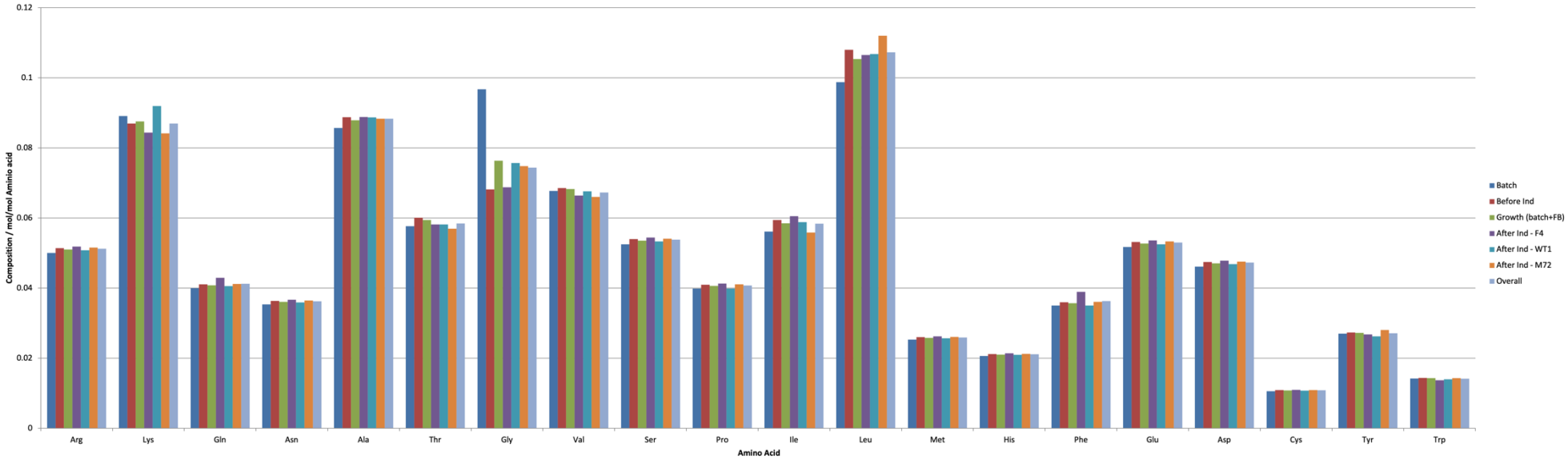

Figure 7.5 Average amino acids composition of biomass proteins (mol/mol Amino acids)

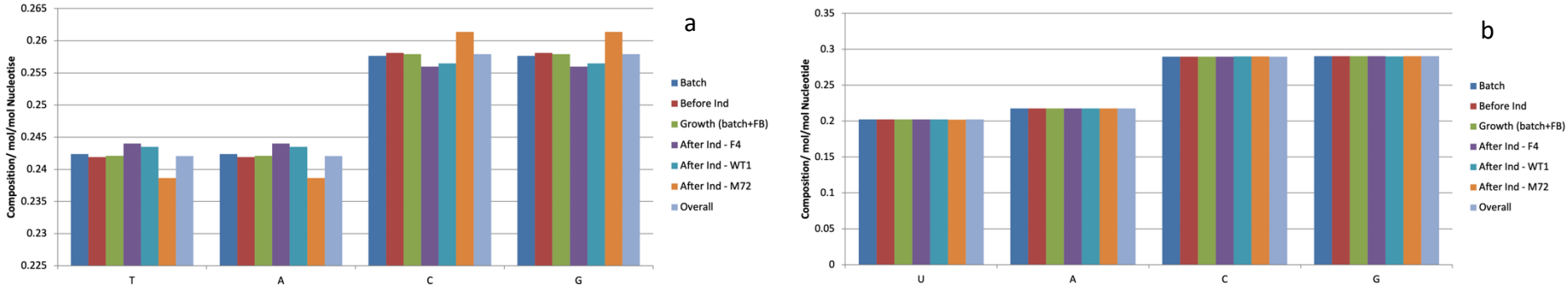

Figure 7.6 Average nucleotide composition for a) DNA and b) RNA

7.3.3 Respiratory Quotient

Calculation of the respiratory quotient (RQ) allows an indirect, but quick, way of identifying the lack of substrate in the growth medium [150] [151]. Equation (7.9) is used to estimate the RQ with OUR representing the oxygen uptake rate and CPR, the carbon dioxide production rate.

$$RQ = \frac{CPR}{OUR} \tag{7.9}$$

For the process data, the RQ's were calculated and can been seen in Figure 7.7. Fermentations MME17 and MME18 had large errors in the off-gas data due to measurement issues. Therefore, their corresponding RQ's have not been calculated.

Figure 7.7 shows that a steady RQ of 1 is achieved for all fermentations after the fed batch process has been initialised. If growth is fully respiratory and the main source is a carbohydrate, in this case glucose, the RQ will be ~1 [152].

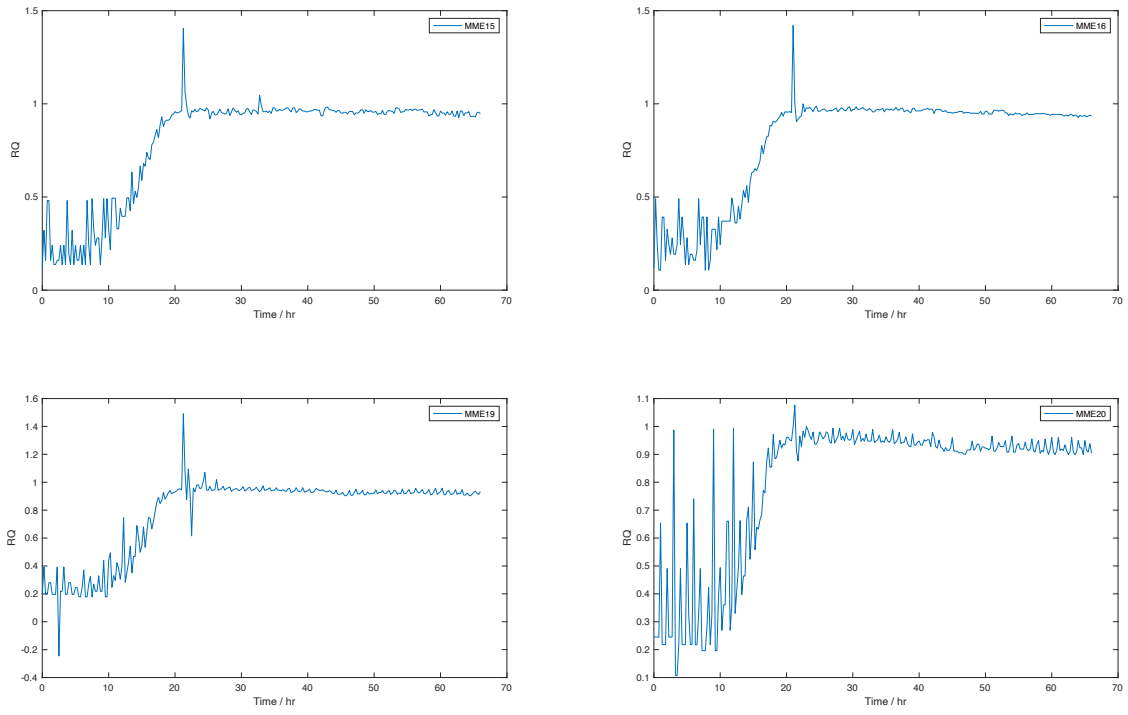

Figure 7.7 RQ for MME15, MME16, MME19 and MME20 against time

7.4 Gene Data

The regulation of genes allows a cell to respond to environmental changes as it changes the set of proteins present. GSK provided four sets of gene expression data for fermentations MME15 to MME18. There exist 133 genes in the data sets. Gene expression has been normalised to show the relative change from the initial gene concentration, G_i, equation (7.10).

$$\Delta G = \frac{G - G_i}{G_i} \tag{7.10}$$

Gene change figures can be found in the Appendix A. These figures show that the gene expression is affected by the antigen induced, and therefore, produced by the culture. Large relative change in gene expression can highlight reactions that are being used by the cell. In the *E. coli* core out of 136 genes, 120 are expressed by one reaction. Knowledge of which reactions are being used the most by the cell would allow for a huge reduction in the problem size in finding elementary flux modes (EFMs). For example, in MME15 and MME16 *amtB* has a peak relative change. This gene is expressed solely by the reversible ammonia transport reaction. When finding EFMs this could be utilised to reduce the search space. Due to the time required to execute this analysis, this work has not been done in this thesis but will be discussed in future work. Whereas *ackA* has a low relative change, so the acetate kinase reaction that expresses it is effectively inactive.

Overall, the gene expressions for same antigen producing cultures are similar. Due to the low relative change of some genes, it would be possible to use these results to reduce the search space for EFMs. This is discussed further in Chapter 7.

7.5 Flux Analysis

Flux analysis of the *E. coli* genome would allow a metabolic map of product formation and substrate consumption to be generated. Metabolic flux analysis (MFA) requires measured fluxes to approximate the unknowns, Chapter 4. Extracellular metabolites, and therefore reactions that cross the cell boundary, must have known concentrations to be able to compute the MFA result [153]. The core network contains 20 extracellular metabolites and only 7 of these are measured in the GSK data set. These metabolites are beta-D-Glucose, ammonia, acetate, L-glutamine, L-glutamate, oxygen, and carbon dioxide. Zamorano *et al* tested the accuracy of underdetermined flux results from MFA on a Chinese hamster ovary cell network [154]. This required ensuring the mass balance system was well-posed to reduce flux intervals via assumptions. They found that although some fluxes were uniquely determinable, the size of certain flux intervals could not be reduced due to the existence of parallel linking between pathways. Therefore, an accurate MFA is not possible for this data set.

Flux balance analysis (FBA) can be used to maximise an objective to find the unknown fluxes, section 3.2, and works with underdetermined systems [153]. This technique will help build a

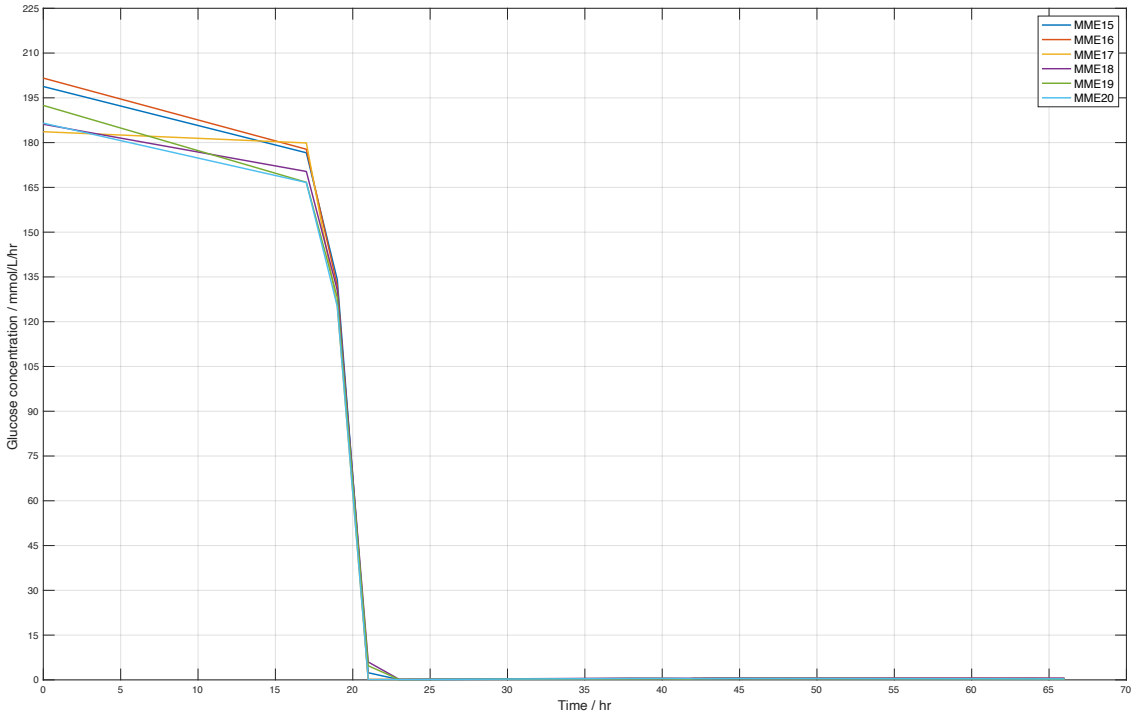

Figure 7.8 Glucose concentration across the experimental time for all fermentations

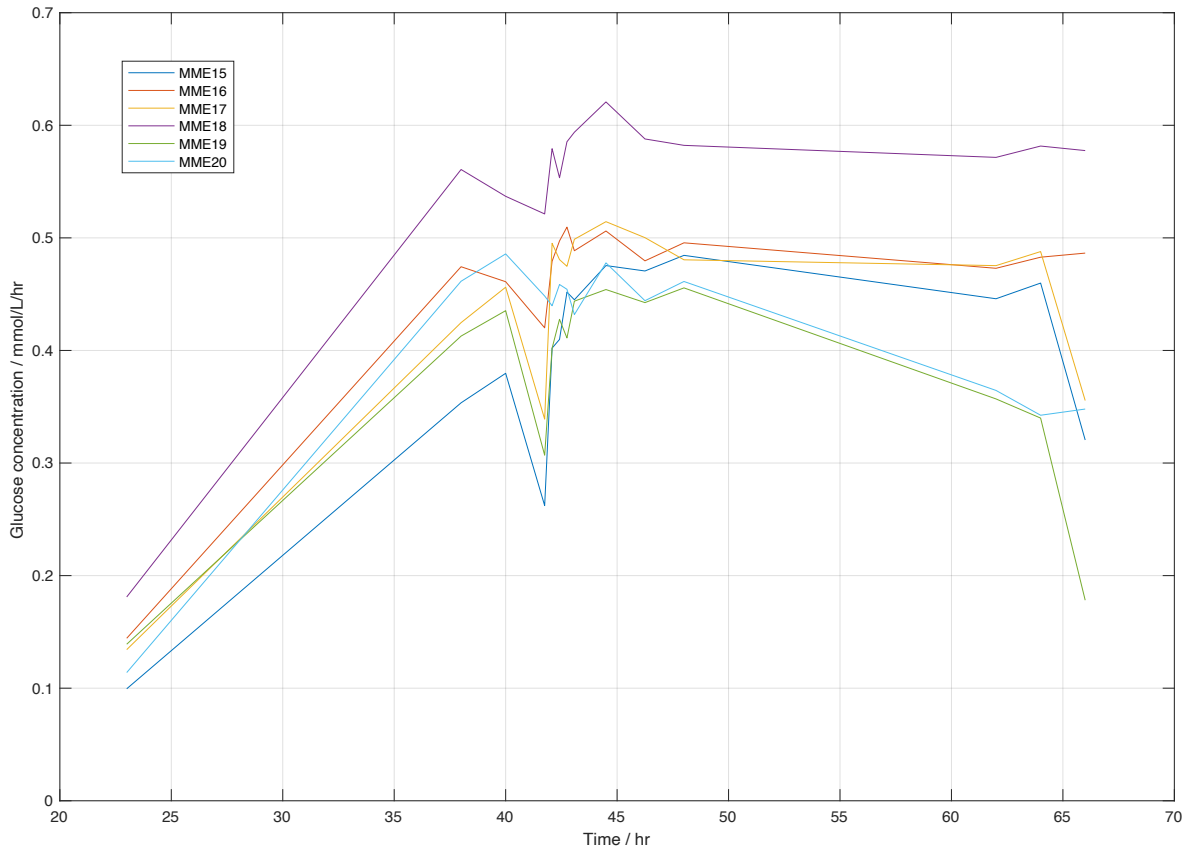

Figure 7.9 Fed-batch glucose concentration in all fermentations

predictive model for a desired optimisation [78]. Due to the complexity of the genome, and the lack of measurable data the data provided was applied to the *E. coli* core network. Maximising biomass is the desired objective which does require the glucose uptake rate to be calculated. The experimental data for glucose concentration is shown in Figure 7.8, with the fed-batch process expanded in Figure 7.9.

7.5.1 Prediction of Biomass Growth in Batch Phase

Biomass growth during batch phase is exponential, equation (7.11). The predicted values for a and b are given in Table 7.1. All R^2 values are very close to, or exactly 1. Therefore, the exponential fit is a good prediction for biomass concentration over time. The average equation for all fermentations is given in equation (7.12) and is shown with experimental data on Figure 7.10.

$$f(X) = a \cdot e^{bt} \tag{7.11}$$

$$f(X) = 0.00011667 \cdot e^{0.55955t} \tag{7.12}$$

Table 7.1 Exponential fitted equation and R2 value for each fermentation

Fermentation	a	b	R^2
MME15	0.0001	0.5574	0.99984
MME16	0.0002	0.5258	0.99946
MME17	0.0001	0.5774	0.9992
MME18	0.0001	0.5643	1.0000
MME19	0.0001	0.5715	0.9999
MME20	0.0001	0.5609	0.9999

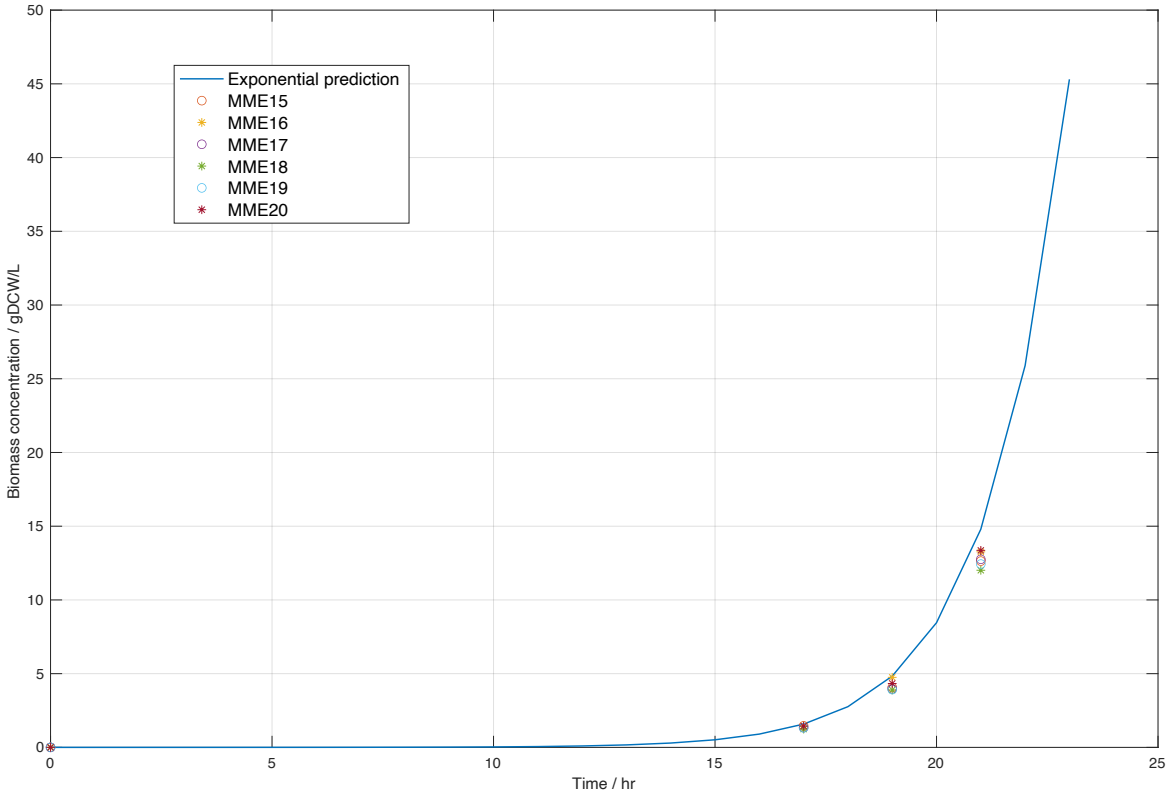

Figure 7.10 Exponential prediction of biomass growth during batch phase

The R^2 values ranged from 0.90003 for MME18 to 0.97841 for MME20. The residuals and the biomass prediction for these 'worst and 'best' cases are given in Figure 7.11 and Figure 7.12 respectively. Due to the good R^2 values achieved, equation (7.12) is a suitable estimation for batch concentration during batch phase for these process conditions.

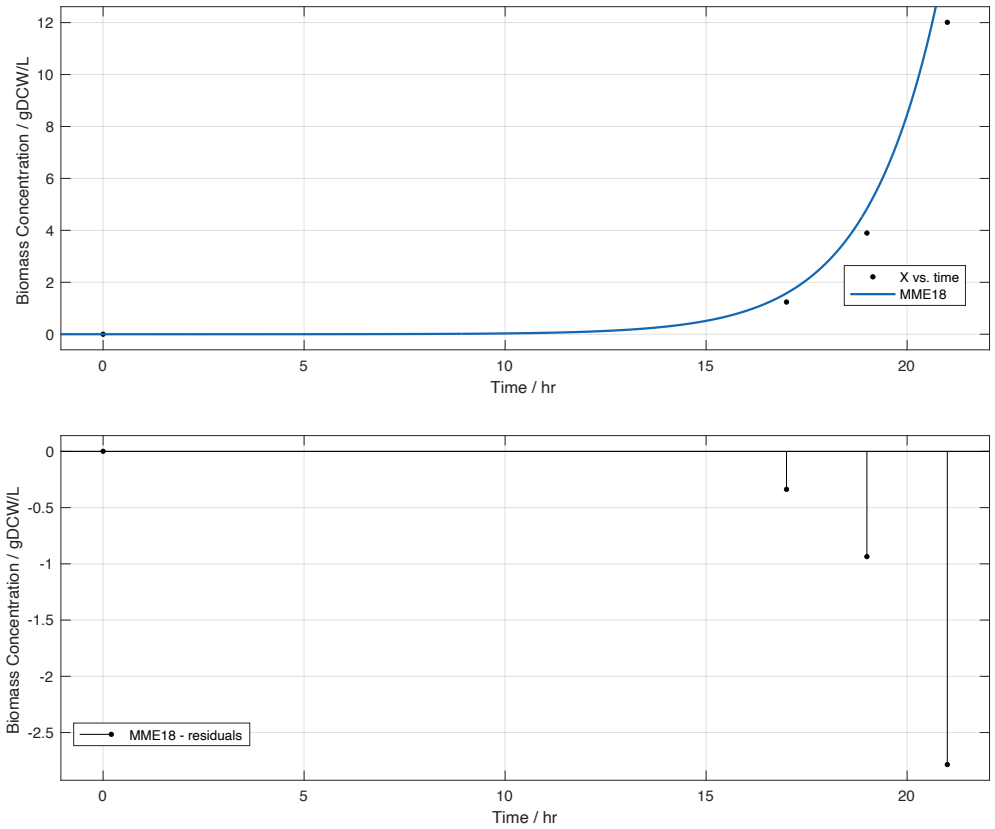

Figure 7.11 'Worst' fermentation prediction case for MME18 with residual plot

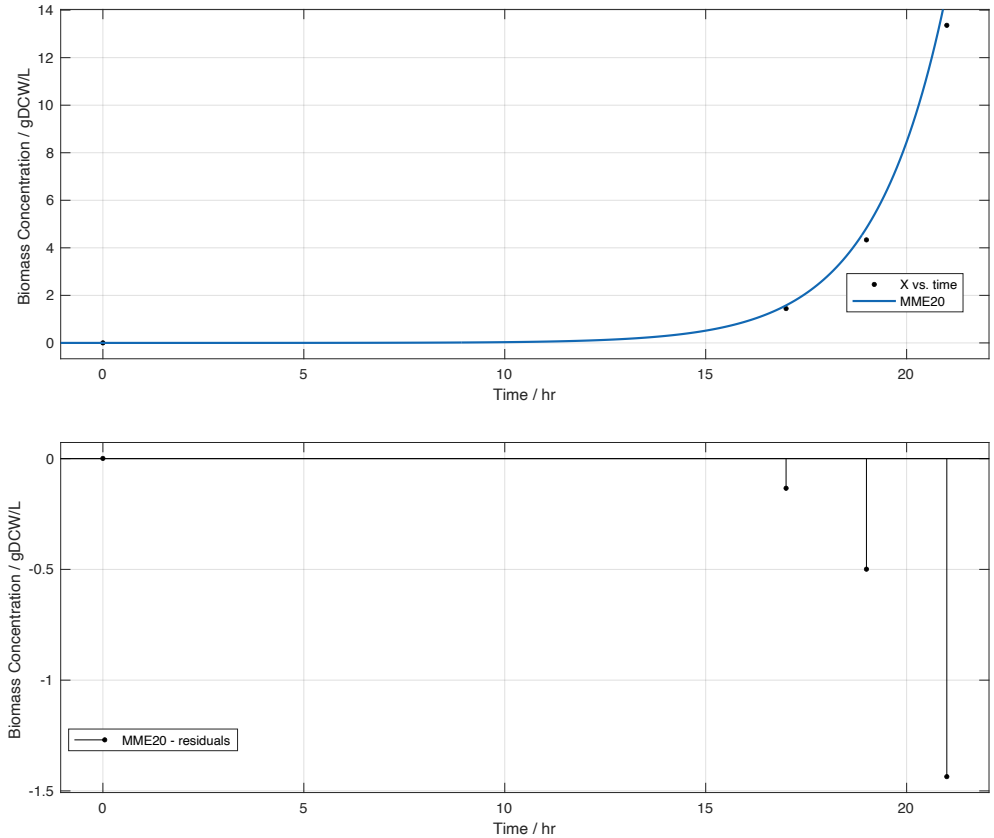

Figure 7.12 'Best' fermentation prediction case for MME20 with residual plot

7.5.2 Prediction of Biomass Growth in Fed-Batch Phase

Biomass experimental data was used to create an equation to predict the growth of the cell during the fed-batch phase. It was found that 3-degree polynomial gave the best fit (provided R^2 close to 1) across all the fermentations. An example of the polynomial fitted to the data for MME15 is given in Figure 7.13.

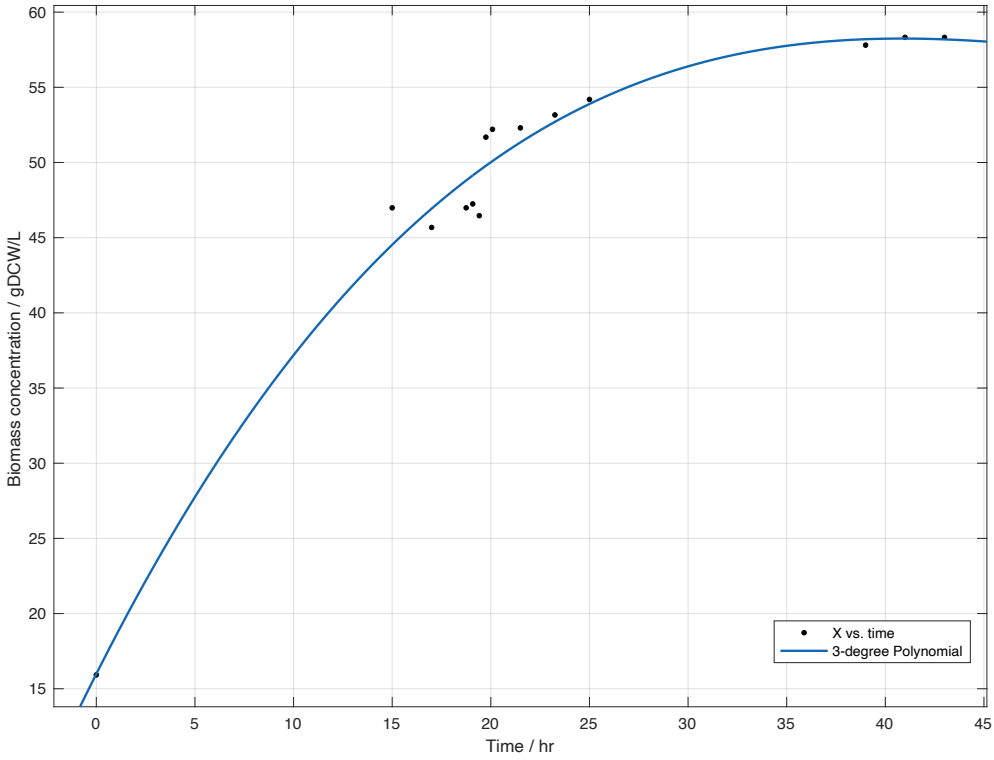

Figure 7.13 Polynomial fitting of MME15 biomass data over time

The fitted equation in the form given in Equation (7.13), along with the R^2 value, is given in Table 7.2.

$$f(X) = P_1 t^3 + P_2 t^2 + P_3 t + P_4 \tag{7.13}$$

Table 7.2 Polynomial fitted equation and R^2 value for each fermentation

Fermentation	P_1	P_2	P_3	P_4	R^2
MME15	0.0003	-0.0514	2.6024	15.9558	0.97719
MME16	0.0002	-0.0406	2.2753	16.4781	0.99293
MME17	0.0003	-0.0419	2.1646	15.1519	0.98791
MME18	0.0002	-0.0405	2.2677	14.7384	0.98549
MME19	0.0003	-0.0527	2.5555	14.2367	0.99413
MME20	0.0006	-0.0752	2.8756	15.4146	0.99914

The average of all these values were taken to generate an overall equation to predict the growth of biomass in fed-batch operation in the process conditions presented in section 7.3, equation (7.14). Figure 7.14 shows equation (7.14) along with all fermentation biomass concentrations over time.

$$X(t) = 0.00031667t^3 - 0.0503833t^2 + 2.45685t + 15.32925 \qquad (7.14)$$

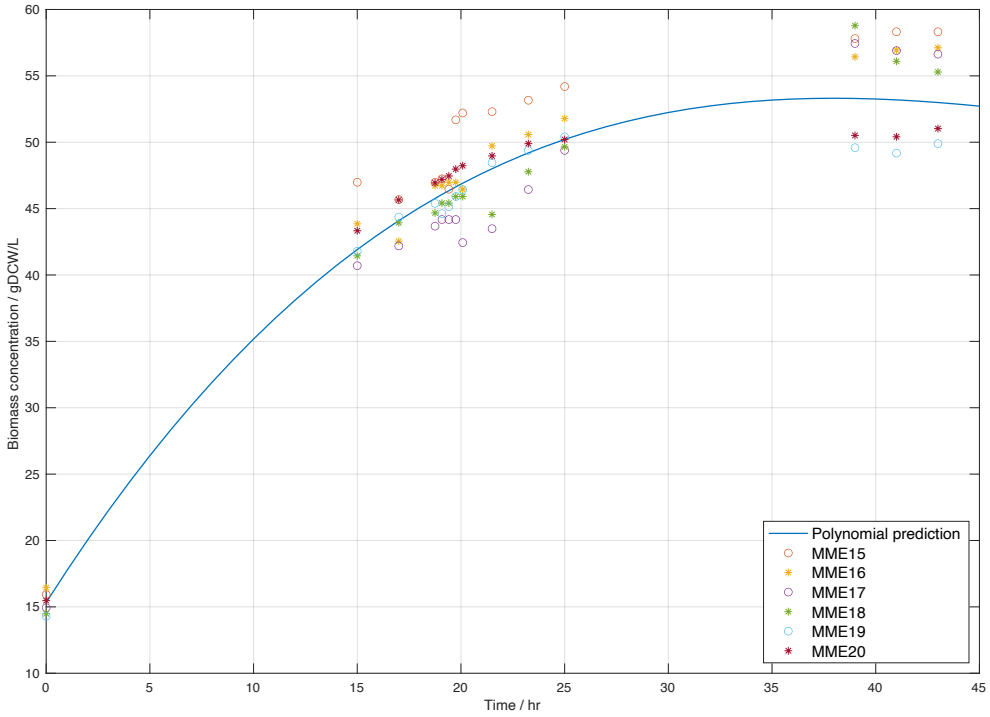

Figure 7.14 Average polynomial for biomass growth during fed-batch operation with all fermentation data points

The R^2 values achieved for the fitted equation range from 0.85483, for MME15, to 0.96918, for MME20. The fit along with the residuals for the 'best' and 'worse' fit are given in Figure 7.16 and Figure 7.15 respectively. Due to the high R^2 values across the 6 fermentations it is fair to assume that the equation (7.14) offers a good prediction of biomass concentration during fed-batch operation.

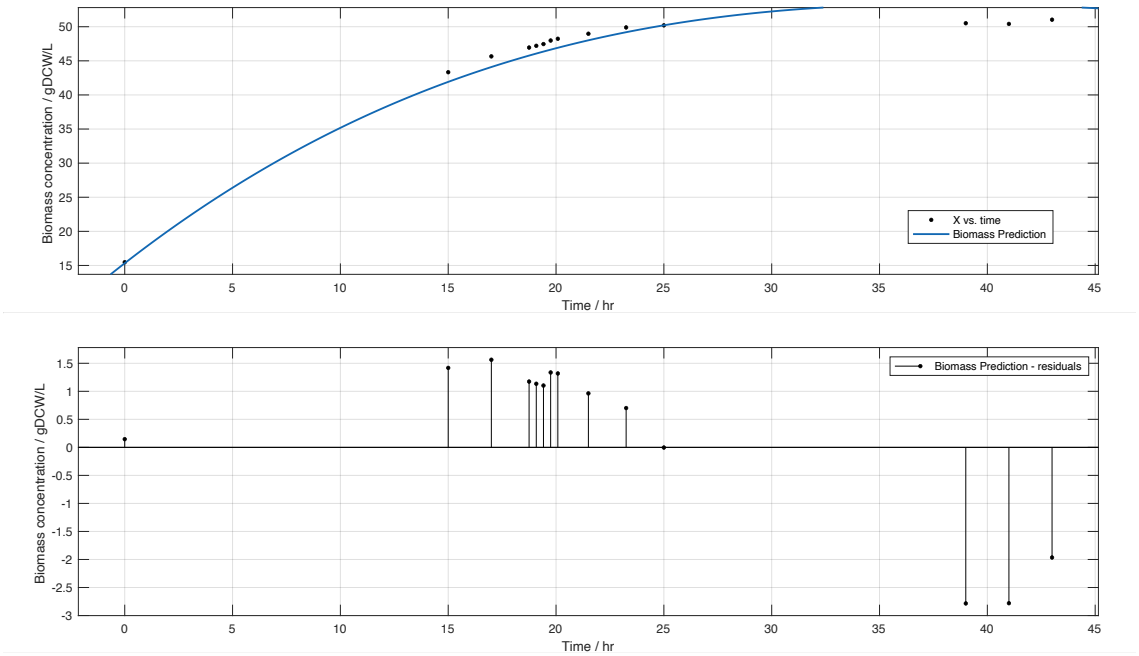

Figure 7.16 MME20 fit with biomass polynomial equation and residuals

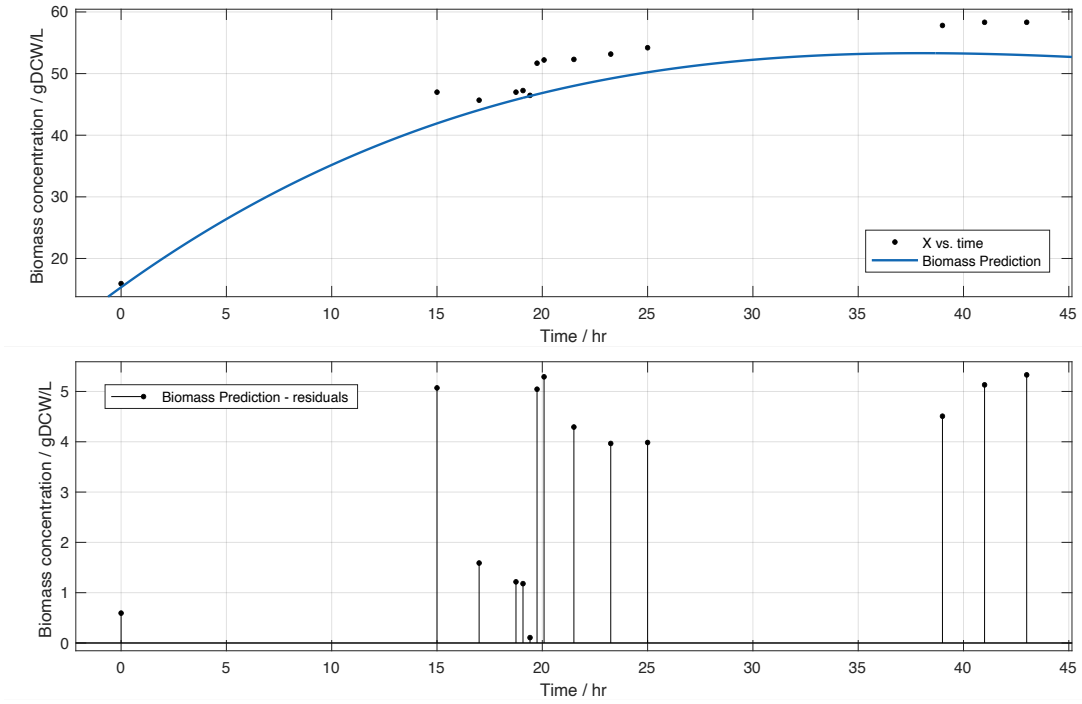

Figure 7.15 MME15 fit with biomass equation and residuals

7.5.3 Specific Growth Rate of Biomass during Batch Operations

The general equation presented in section 7.5.1 is one way of predicted biomass growth. During the batch phase of a cell's life, the biomass concentration can also be modelled by equation (7.15), where μ is the specific growth rate of biomass in gDCW L^{-1} hr^{-1}. Integration of equation (7.15) yields equation (7.16), which with the application of least squares allows μ to be found, equation (7.17).

$$\frac{dX}{dt} = \mu X \tag{7.15}$$

$$ln\frac{X}{X_0} = \mu t \tag{7.16}$$

$$\mu = \frac{\sum ln\frac{X}{X_0}}{\sum t} \tag{7.17}$$

The estimated μ for each fermentation is given in Table 7.3. The R^2 values for all fermentations are close to 1, therefore, proving the calculated μ offer good biomass concentration prediction in the batch phase. This is further emphasised by Figure 7.17, which shows the actual vs predicted response for all fermentations. The similar μ achieved for each fermentation shows that the biomass growth rate is not antigen dependent.

Table 7.3 Estimated specific growth rate of biomass in batch phase with the corresponding R^2 number

Fermentation	μ / gDCW L^{-1} hr^{-1}	R^2
MME15	0.4318	0.9269
MME16	0.4311	0.9354
MME17	0.4327	0.9033
MME18	0.4304	0.9052
MME19	0.4298	0.9056
MME20	0.4342	0.9184

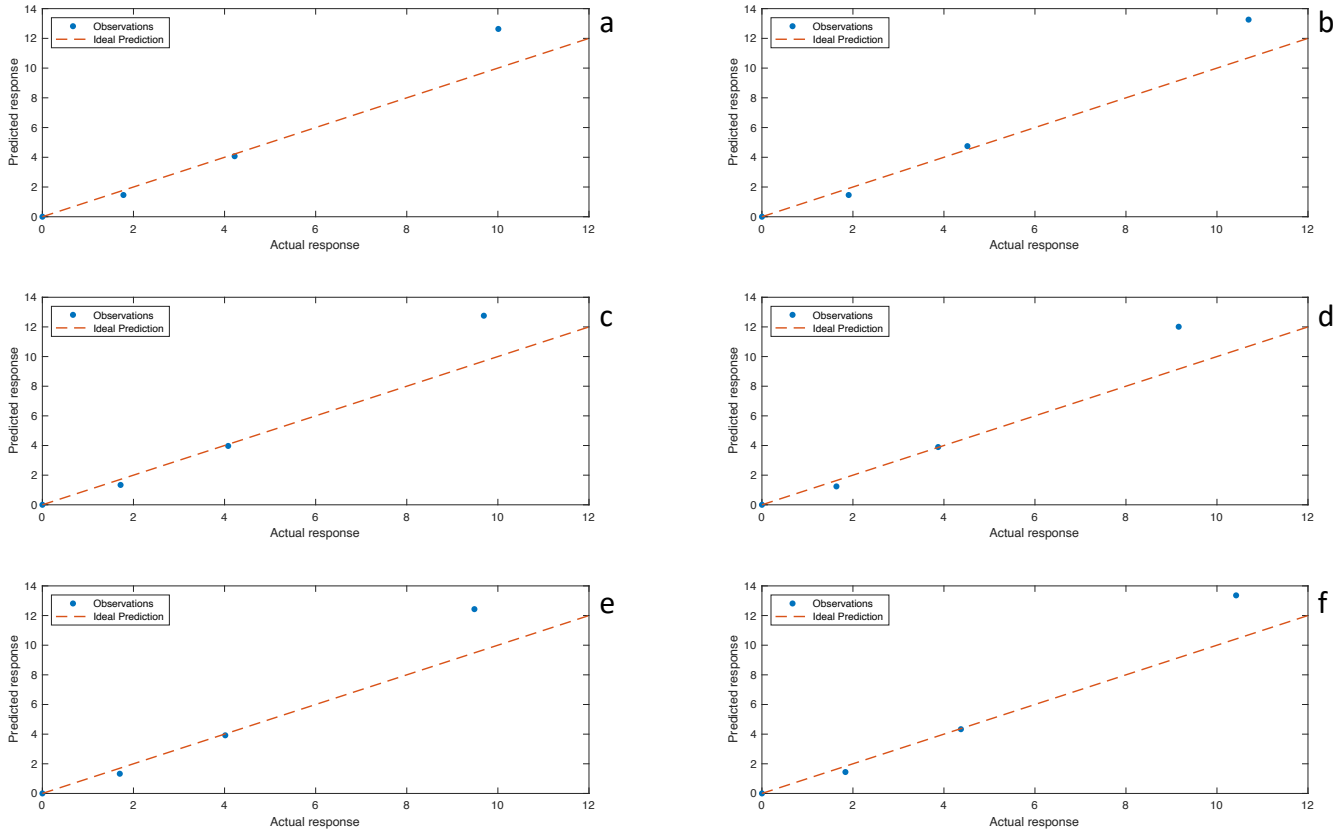

Figure 7.17 Actual vs predicted response of biomass concentration using calculated specific growth rates for a) MME15 b) MME16 c) MME17 d) MME18 e) MME19 f) MME20

7.5.4 Specific Uptake Rate of Glucose during Batch Operations

In batch operations the change in substrate concentration is determined via equation (7.18), where v is the specific uptake rate in mmol L^{-1} hr^{-1}. Combining this with equation (7.15) generates a, relating specific biomass growth and specific uptake rate, equation (7.19). Integration of equation (7.18) allows for a to be calculated with least squares, equation (7.20).

$$\frac{dS}{dt} = -vX \tag{7.18}$$

$$\frac{dS}{dX} = -\frac{v}{\mu} = -a \tag{7.19}$$

$$a = -\frac{\sum S - S0}{\sum X - X0} \tag{7.20}$$

Glucose consumption was simulated, and the simulation output compared to the experimental data. The specific biomass growth rate for each fermentation was also used to generate the simulation.

Table 7.4 provides the calculated specific uptake rates of glucose for each fermentation in batch phase. The R^2 values achieved show that the estimated specific uptake rates are realistic estimations.

Table 7.4 Estimated specific uptake rate of glucose in batch phase with the corresponding R^2 number

Fermentation	a **/ mmol**	v **/ mmol gDCW^{-1} hr^{-1}**	**R^2**
MME15	15.5979	6.7349	0.9339
MME16	15.1499	6.5310	0.9398
MME17	13.6040	5.8863	0.8670
MME18	14.6728	6.3152	0.8975
MME19	15.7187	6.7556	0.9934
MME20	13.9900	6.0741	0.9253

Figure 7.18 shows the actual response and the predicted response for glucose concentration using estimated biomass growth rate and glucose uptake rate. The fit achieved is good until about 21 hours. At this point nearly all glucose has been consumed by the cells, Figure 7.8, changing how the cell is behaving. Equations (7.14) and (7.17) only account for batch growth phase with a good supply of growth medium. Therefore, it is expected that at 21 hours the estimated rates will not encapsulate the cell's behaviour. However, the R^2 values show that the estimations are still decent enough to model glucose consumption with the estimated specific uptake rates in the batch phase.

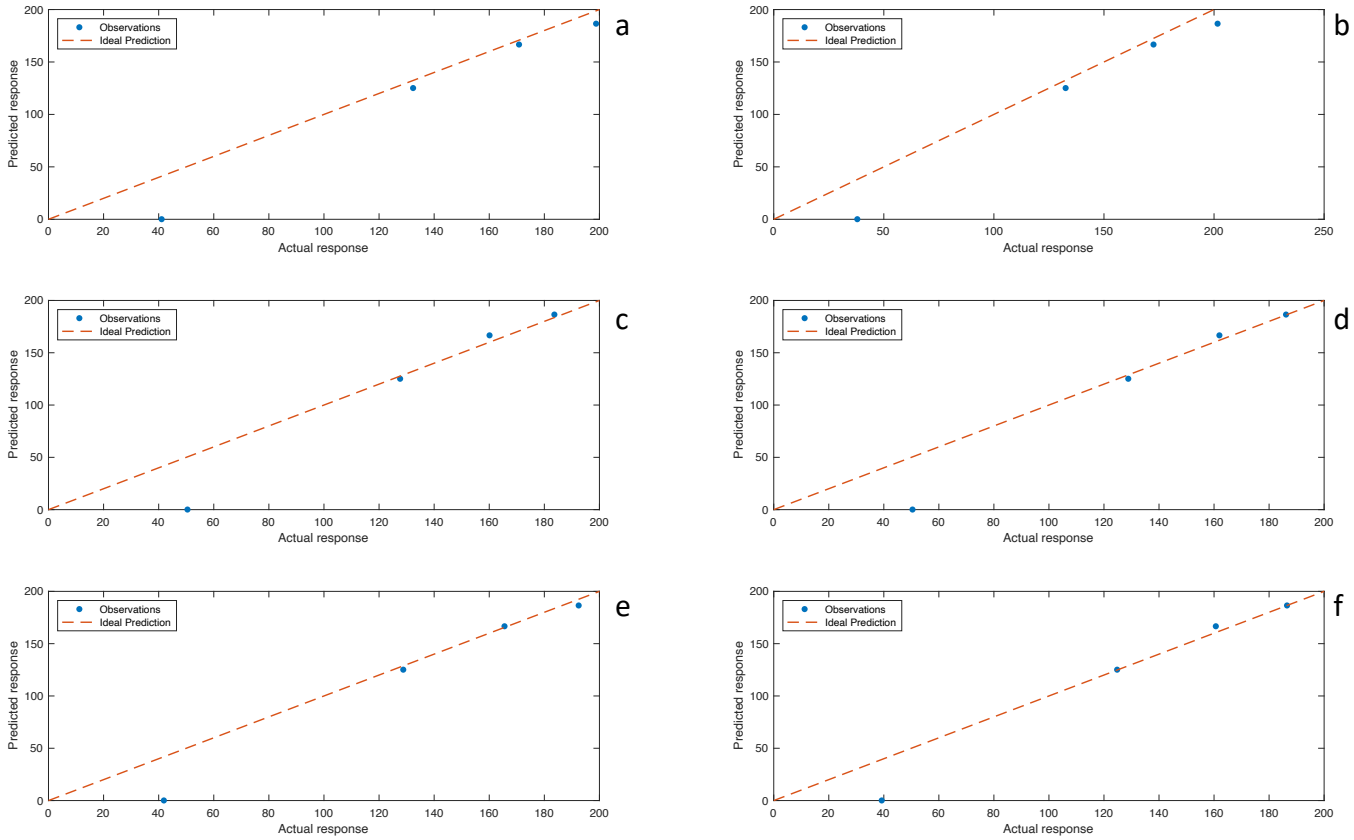

Figure 7.18 Actual vs predicted response of glucose concentration using calculated specific growth rates for a) MME15 b) MME16 c) MME17 d) MME18 e) MME19 f) MME20

7.5.5 Specific Growth Rate of Biomass during Fed-Batch Operations

The system, although fed batch, operates as mini batch systems as the feed is not continuous. Each 15-minute interval is a batch process. To reduce the amount of data created it is assumed that each hour is a batch process. Figure 7.19 shows the biomass concentration over the fed batch period along with the predicted biomass estimated using μ from equation (7.16). The figure shows the excellent fit achieved by modelling the system as batch over each hour. This is not just the case for MME15; good fits were achieved for each fermentation as shown in Figure 7.20. The specific growth rate is greatest initially and reduces over the experimentation time as the cell enters the transition and decay phases of life. The growth rates for each hour for all fermentations are shown in Figure 7.21. During the batch phase the specific growth rate was around 0.4gDCWL^{-1}h^{-1}, therefore, initial average rates of 0.1146gDCWL^{-1}h^{-1} are reasonable.

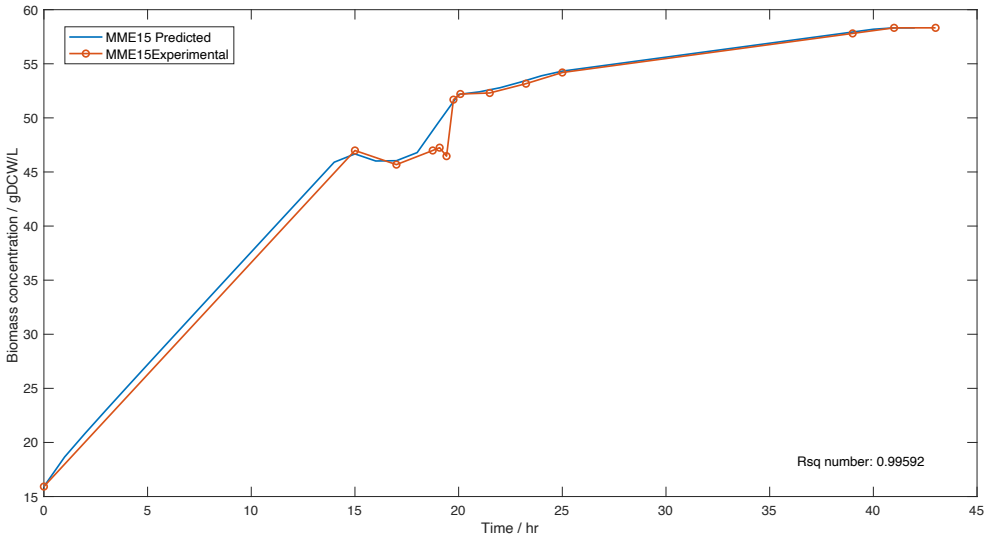

Figure 7.19 MME15 biomass experimental data and predicted data using estimated specific growth rate

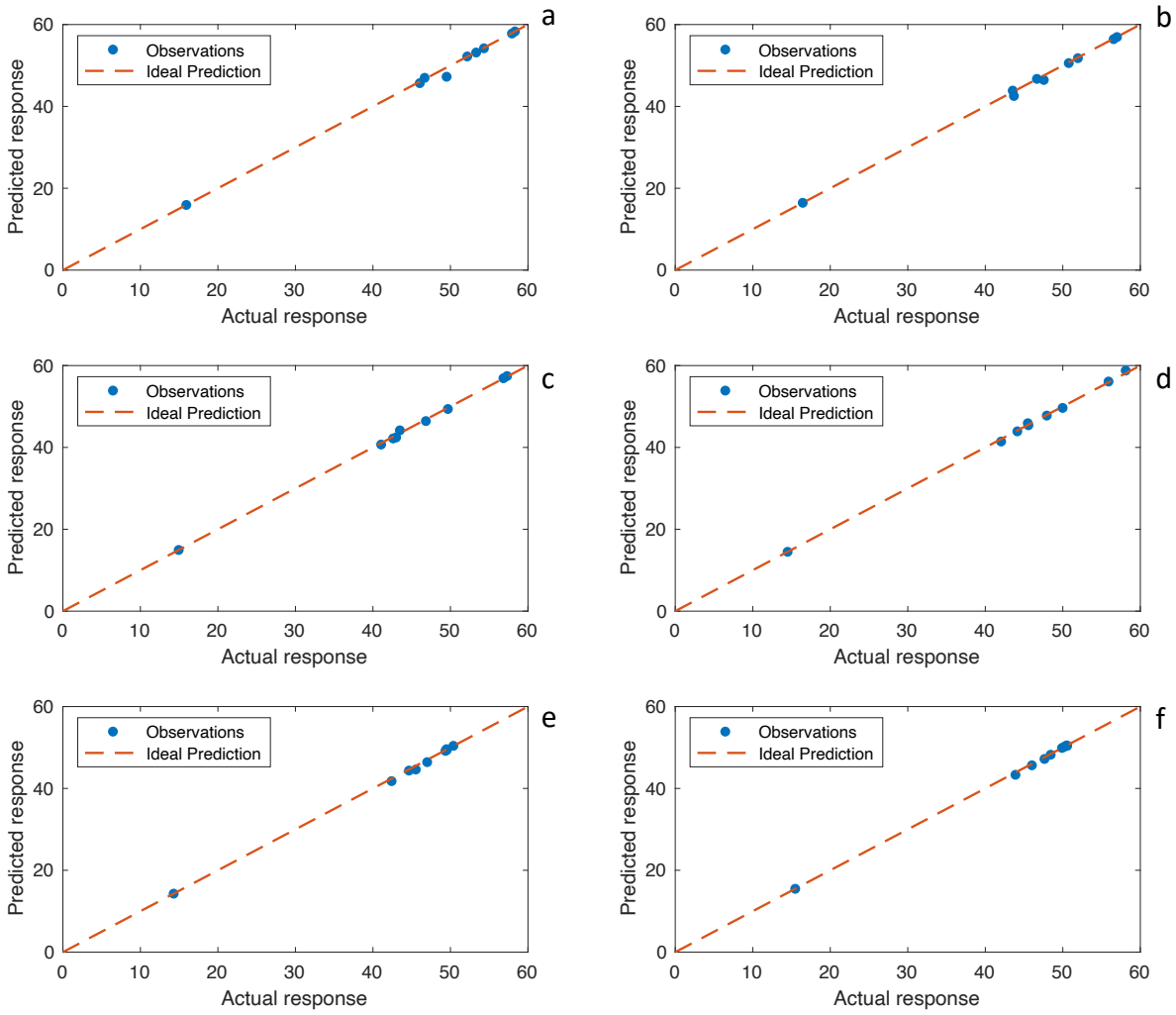

Figure 7.20 Actual vs predicted response of biomass concentration using calculated specific growth rates for a) MME15 b) MME16 c) MME17 d) MME18 e) MME19 f) MME20

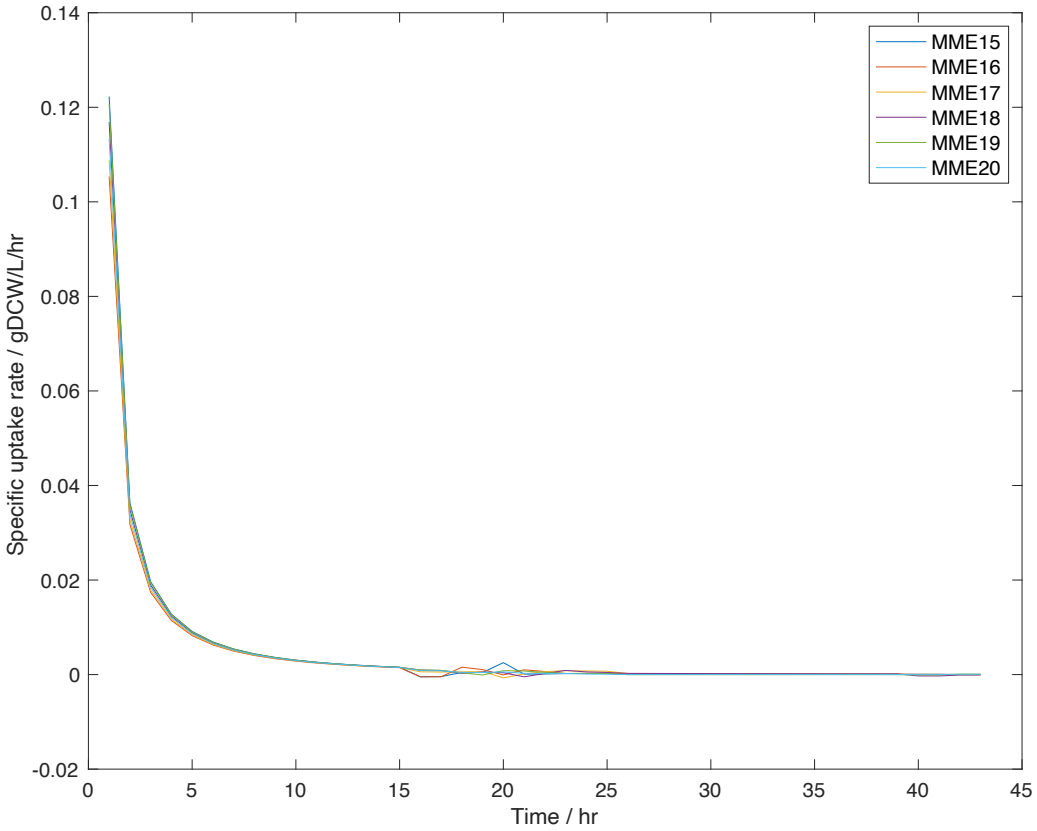

Figure 7.21 Specific uptake rates over fed batch for all fermentations

7.5.6 Specific Uptake Rate of Glucose during Fed-Batch Operations

Like in section 7.5.5 the fed batch system can be multiple batch sections. During the fed batch process a control scheme is utilised to feed glucose to the cells. This control scheme aims at maintaining the glucose concentration so that $\frac{d(Glucose)}{dt} = 0$. As the change in glucose concentration should be 0 it can be assumed that the glucose feed is equivalent to the glucose uptake rate,

$$if\ \frac{d(Glucose)}{dt} = 0 \therefore v = c_F \tag{7.21}$$

The glucose feed rate is measured for all fermentations, therefore, based on equation (7.21) the uptake rate is also known. However, as can be seen by Figure 7.9 the control scheme does not achieve $\frac{d(Glucose)}{dt} = 0$ across the fed batch phase,

$$if\ \frac{d(Glucose)}{dt} > 0 \therefore v < c_F \tag{7.22}$$

$$if\ \frac{d(Glucose)}{dt} < 0 \therefore v > c_F \tag{7.23}$$

Due to the small changes in concentration over time the increase or decrease in v compared to c_F will also be small. This will only marginally affect the fluxes achieved by FBA, but the on/off state of reactions will still be clear. Overall, this work assumes an ideal control scheme, so equation (7.21) is upheld. The feed concentration, equivalent to uptake rate, is given in Figure 7.22. The negative feed concentration at 1 hour for MME20 is due to a measurement error caused by the mass measurement of the feed vessel every 15 minutes.

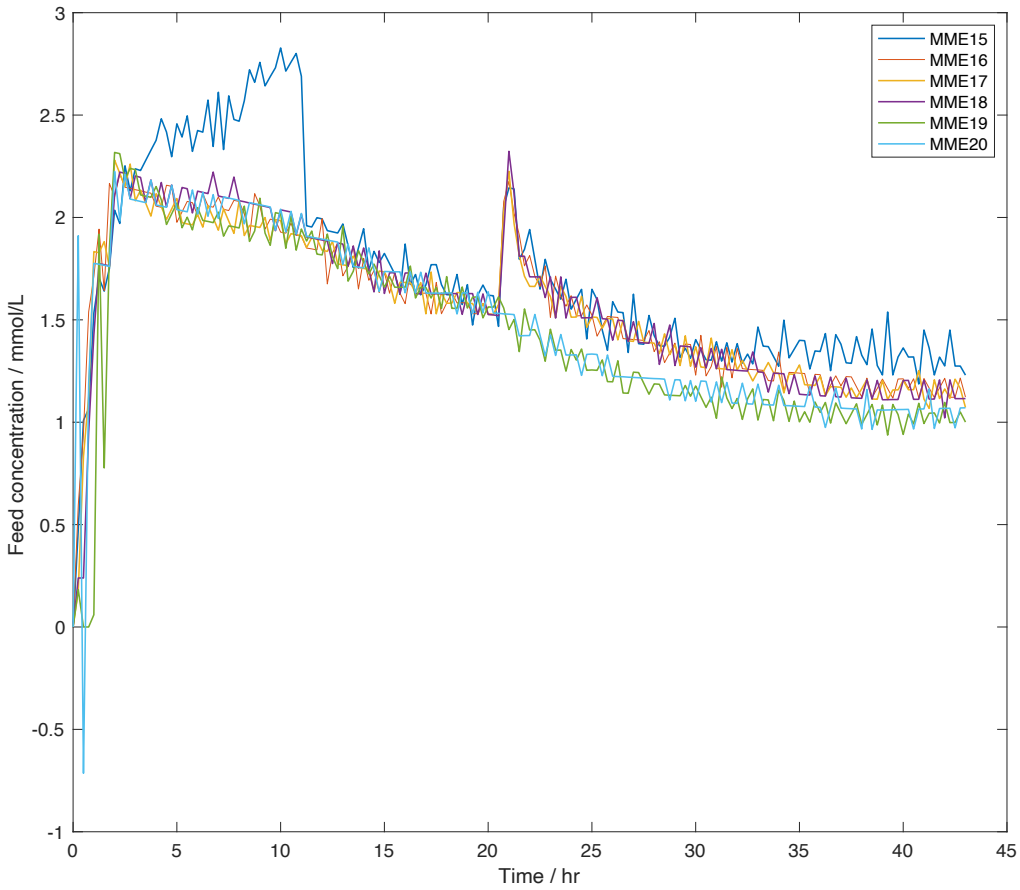

Figure 7.22 Glucose feed concentrations over the fed batch phase

Rather than utilising the specific uptake rate every 15 minutes, the mean uptake rate will be used for flux analysis in the fed batch phase. Table 7.5 details the average uptake rate of glucose for each fermentation in the fed batch phase. These results are less than that during

the batch phase, which is expected as the cell is leaving the growth phase as fed batch is initiated.

Table 7.5 Estimated specific uptake rate of glucose in fed batch phase

Fermentation	v / mmol gDCW^{-1} hr^{-1}
MME15	1.4832
MME16	1.4325
MME17	1.5695
MME18	1.5554
MME19	1.5730
MME20	1.7198

7.6 Flux Balance Analysis

FBA can be efficiently performed on the COBRA Toolbox, which has the *E. coli* core network readily available. The default constraints on the toolbox are ATP maintenance requirement lower bound is 8.39 mmol gDCW^{-1} hr^{-1} and the D-Glucose exchange reaction lower bound is -10 mmol gDCW^{-1} hr^{-1}. The D-Glucose lower bound is adjusted to reflect the uptake rates found for the batch and fed-batch operations. As the process is aerobic, high oxygen uptake rates are required. During the batch operations the oxygen uptake rate peak is approximately 1700 mmol hr^{-r} and 1000 mmol hr^{-1} in fed-batch. The biomass equation, which is maximised, is given in Table 7.6. Biomass formation is given in the commonly used Neidhardt biomass equation with growth-associated maintenance [38, 155].

Table 7.6 E. coli core biomass equation

Metabolite	Stoichiometry
3-Phospho-D-glycerate	-1.496
Acetyl-CoA	-3.7478
ADP	59.81
2-Oxoglutarate	4.1182
ATP	-59.81
Coenzyme A	3.7478

D-Erythrose 4-phosphate	-0.361
D-Fructose 6-phosphat	-0.0709
Glyceraldehyde 3-phosphate	-0.129
D-Glucose 6-phosphate	-0.205
L-Glutamine	-0.2557
L-Glutamate	-4.9414
H2O	-59.81
H+	59.81
Nicotinamide adenine dinucleotide	-3.547
Nicotinamide adenine dinucleotide - reduced	3.547
Nicotinamide adenine dinucleotide phosphate	13.0279
Nicotinamide adenine dinucleotide phosphate - reduced	-13.0279
Oxaloacetate	-1.7867
Phosphoenolpyruvate	-0.5191
Phosphate	59.81
Pyruvate	-2.8328
alpha-D-Ribose 5-phosphate	-0.8977

During batch operations the fluxes for all 95 reactions were calculated, with maximising biomass as the objective. The objective function achieved, which is the predicted biomass growth rate for each fermentation is given in Table 7.7. These growth rates are greater than that achieved in the growth rate estimations, Table 7.3. This is expected as FBA produces the maximum possible rate and when compared to the real system, maximum uptake rate will rarely be achieved due to FBA not accounting for environmental factors during fermentation. These factors, such as temperature and pH, could hinder the growth of the cell over time.

Figure 7.23 displays all the fluxes required across the fermentations to maximise biomass growth rate. Out of 95 reactions, some of which are reversible, FBA estimates only 47 are necessary to achieve these high rates. The flux necessary for each reaction is consistent across the fermentations, despite the different antigens being produced. Therefore, the antigens being produced does not affect the reactions necessary to maximise biomass for *E. coli*. This could be due to the antigens being grown on the bacterial cell wall, which in turn does not affect the internal route taken within the metabolic network inside the wall [156].

FBA results often show fluxes in unrealistic conditions. However, due to the similarity in the maximised specific growth rate and the measured rate estimated by the experimental data

the results shown in Figure 7.23 may show an example of what could be happening within the cell.

Table 7.7 Maximised objective function for each fermentation in batch phase and percentage of increase from estimated rate and maximum possible

Fermentation	μ / mmol gDCW hr^{-1}	Percentage of increase between estimated value and maximised objective
MME15	0.5746	25%
MME16	0.5559	22%
MME17	0.4968	13%
MME18	0.5362	20%
MME19	0.5765	25%
MME20	0.5141	16%

The mean specific uptake rates for the fed batch phase were used to maximise the biomass growth rate. Table 7.8 details the maximum specific growth rates for biomass across the 6 fermentations. These rates are lower than that possible in the batch phase, Table 7.7, as the cell is not in the growth phase. However, the cell will still be growing but at a reduced rate, emphasised by the maximum feasible μ.

Table 7.8 Maximised objective function for each fermentation in fed batch phase

Fermentation	μ / mmol gDCW hr^{-1}
MME15	0.0932
MME16	0.0886
MME17	0.1011
MME18	0.0998
MME19	0.1015
MME20	0.1149

Figure 7.24 provides the FBA results for fed batch operation for all fermentations. Only 44 out of the 95 reactions are necessary to achieve the maximum feasible rates in Table 7.8. To ensure that equation (7.20) could be assumed to be upheld throughout the fed batch process

±5% was added to the mean specific uptake rates. The active reactions required maximise the biomass growth rate remained unchanged.

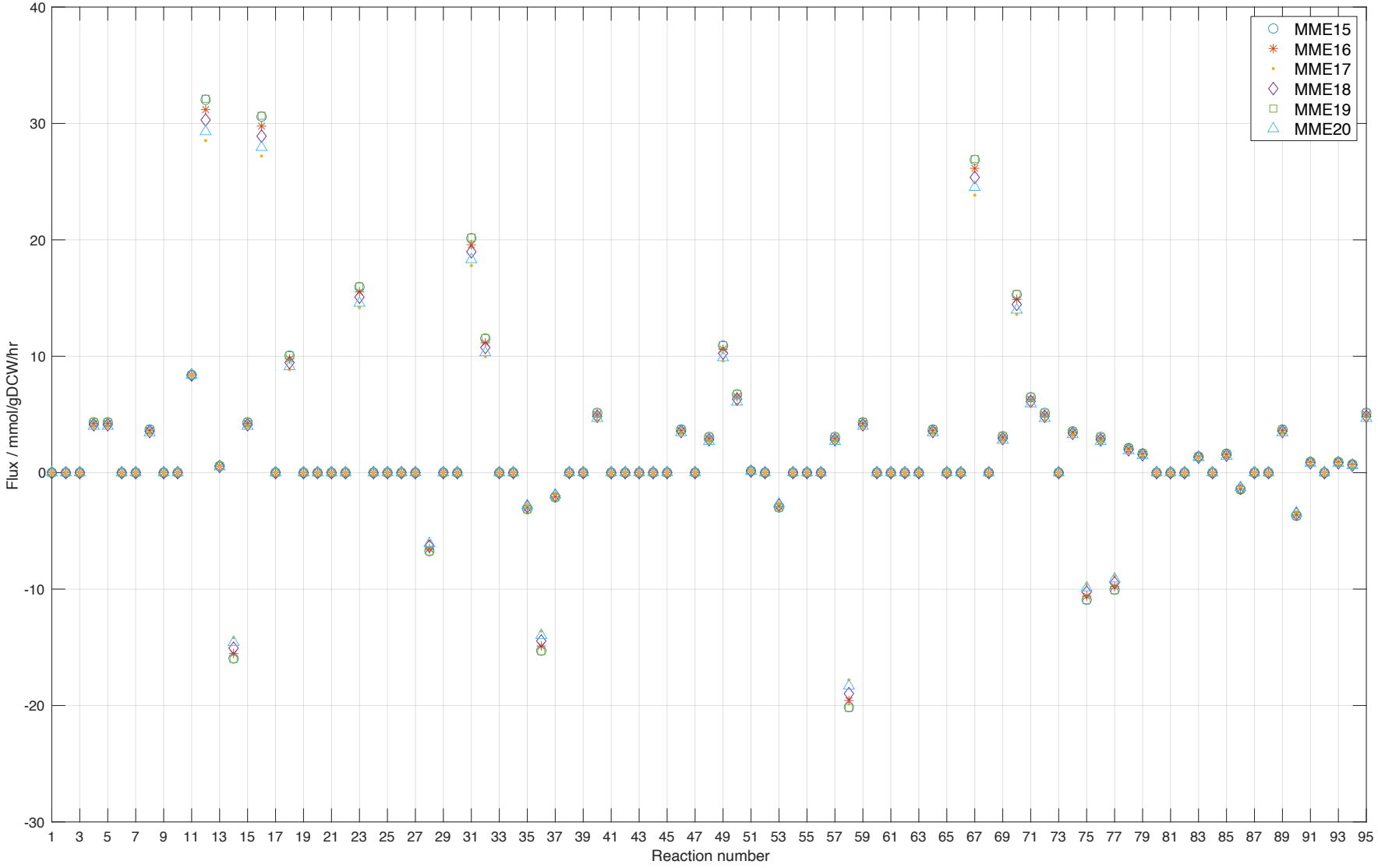

Figure 7.23 Flux for E. coli core network required to maximise biomass growth rate in batch phase

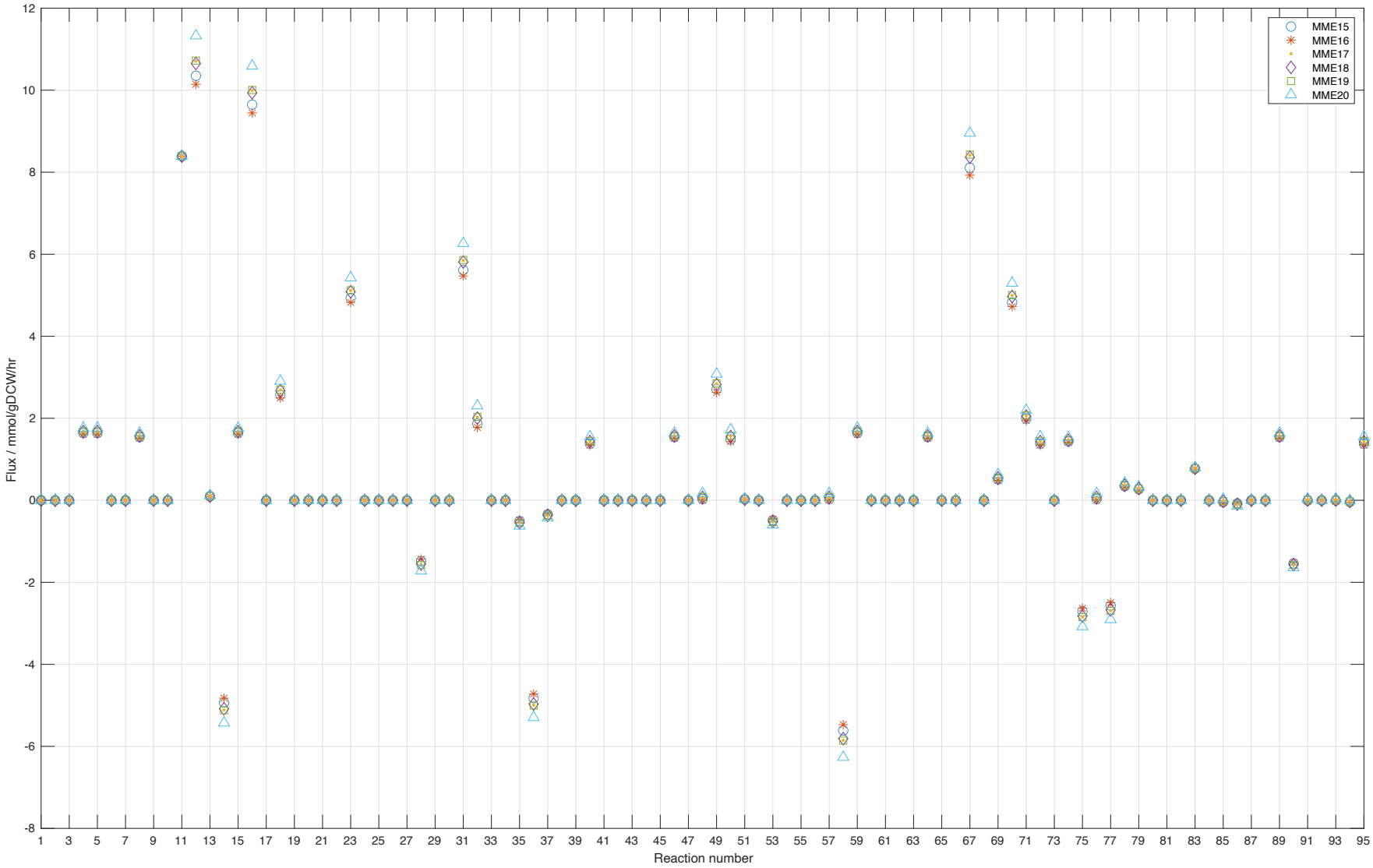

Figure 7.24 Flux for E. coli core network required to maximise biomass growth rate in fed batch phase

Like with the batch phase the fluxes needed are similar across all fermentations, emphasising that the chosen antigen does not affect the active reactions. Three reactions are necessary in the batch system that are not needed in fed batch: reactions 91, 93 and 94. These reactions will be necessary to form the additional EFMs needed in the batch system that are not present in the fed batch solution set. Reactions 91, 93 and 94 are reversible but the flux results from FBA, see Figure 7.23, show that they are all operating in the forward direction. The reactions are,

$$r_{91} \quad Glycerol\text{-}3\text{-}phosphate + Sedoheptulose\text{-}7\text{-}phosphate \rightarrow D\text{-}Erythrose\text{-}4\text{-}phosphate + D\text{-}Fructose\text{-}6\text{-}phosphate$$

$$r_{93} \quad D\text{-}Xylulose\text{-}5\text{-}phosphate + alpha\text{-}D\text{-}Ribose\text{-}5\text{-}phosphate \rightarrow Glycerol\text{-}3\text{-}phosphate + Sedoheptulose\text{-}7\text{-}phosphate$$

$$r_{94} \quad D\text{-}Erythrose\text{-}4\text{-}phosphate + D\text{-}Xylulose\text{-}5\text{-}phosphate \rightarrow Glycerol\text{-}3\text{-}phosphate + D\text{-}Fructose\text{-}6\text{-}phosphate$$

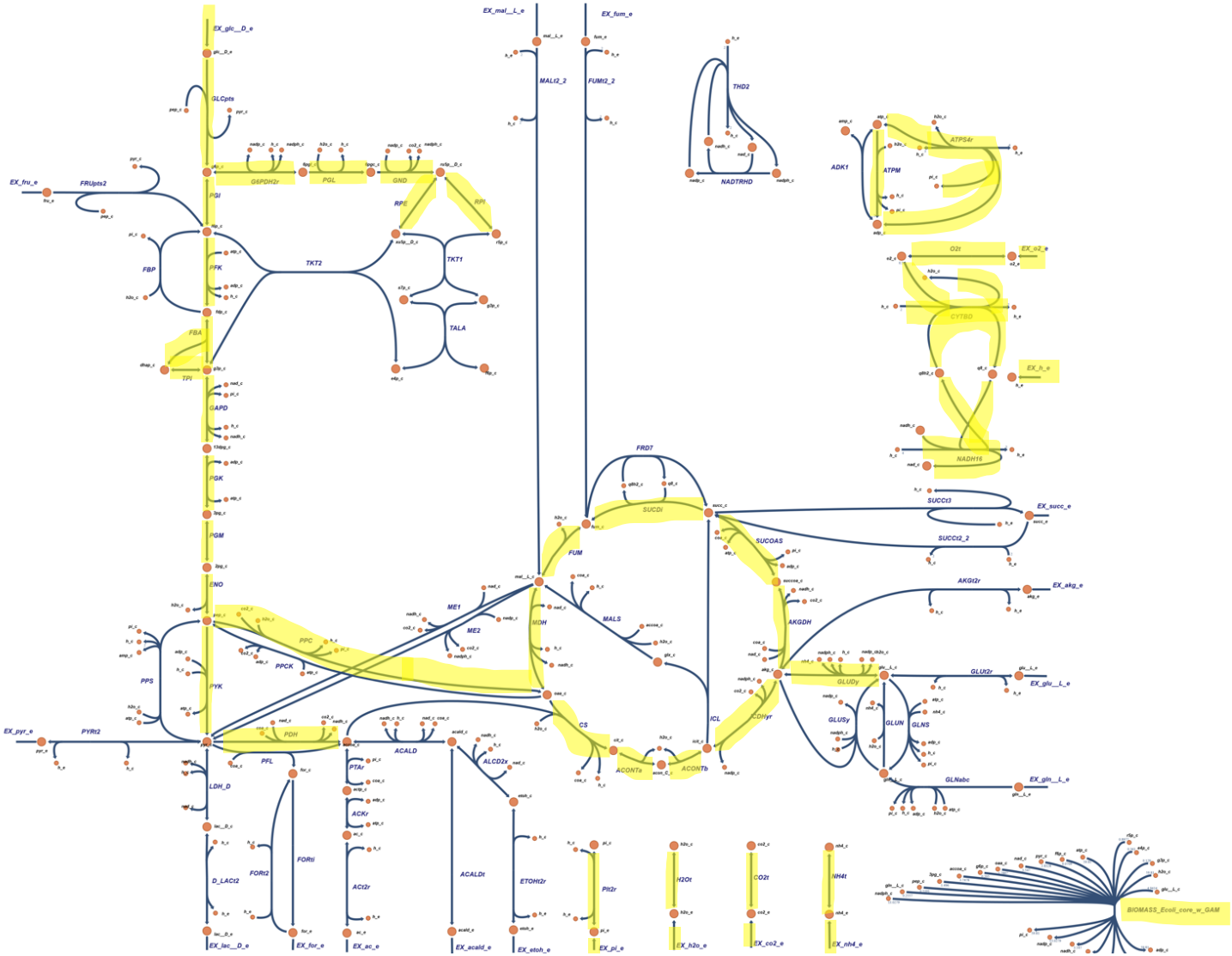

Figure 7.25 Reactions necessary in fed batch phase overlayed onto E. coli core map

The *E. coli* core consists of 95 reactions and 72 metabolites. The EFMs for this network can be found efficiently via efmtool, with the COBRA Toolbox modelling the *E. coli* core [38]. The glucose uptake rate was provided to efmtool for both the batch and fed-batch operations of each fermentation. The batch phase for all fermentations was found to have no EFMs and the fed batch phase 1 EFM. These two phases differ in the number of EFMs as the lower glucose uptake rate and biomass growth rate will make some reactions unfeasible. The reactions necessary in the fed batch phase overlayed onto the *E. coli* core map are shown in Figure 7.25.

Applying flux data to reduce the search space for EFM enumeration will aid in the solve time and memory storage required for MILP. However, FBA presents an ideal non-unique case that does not consider EFMs but instead what reactions are necessary to produce the maximum biomass. As the case is non-unique it is fair to assume that EFMS would occur but instead they are not being encapsulated in the non-unique flux distribution found. A better use of flux analysis in the determination of EFMs would be flux variability analysis (FVA).

7.7 Flux Variability Analysis

FVA was performed around maximising the biomass objective. FVA in the batch phase yields 7 reactions, out of 95 non-decomposed reactions, which have a minimum and maximum flux of 0mmol$L^{-1}hr^{-1}$, Figure 7.26 and Figure 7.27. The same reactions are found to be unused in the fed batch phase, Figure 7.28 and Figure 7.29. The unused reactions are exchange reactions for D-Fructose, fumarate, L-Glutamine and L-Malate, allowing for extracellular metabolites to enter the cell. The other three reactions are,

$$r_{45} \quad D\text{-}Fructose + Phosphoenolpyruvate \rightarrow D\text{-}Fructose\text{-}6\text{-}phosphate + Pyruvate$$

$$r_{52} \quad ATP + L\text{-}Glutamate + Water \rightarrow ADP + L\text{-}Glutamate + Hydrogen + Phosphate$$

$$r_{63} \quad \mathrm{Hydrogen}_e + \mathrm{L\text{-}Malate}_e \rightarrow \mathrm{Hydrogen} + \mathrm{L\text{-}Malate}$$

The used reactions all have minimum and maximum flux ranges shown on the figures. Any reaction can have a flux within this range occurring at any moment in time, whilst still maximising biomass. Some reactions have small ranges allowing for the reasonable assumption that the mean flux is occurring, however, some reactions like reaction 44, have large ranges. Further investigation should be done to estimate the optimal flux necessary to maximise biomass, however further measured data is required to do this.

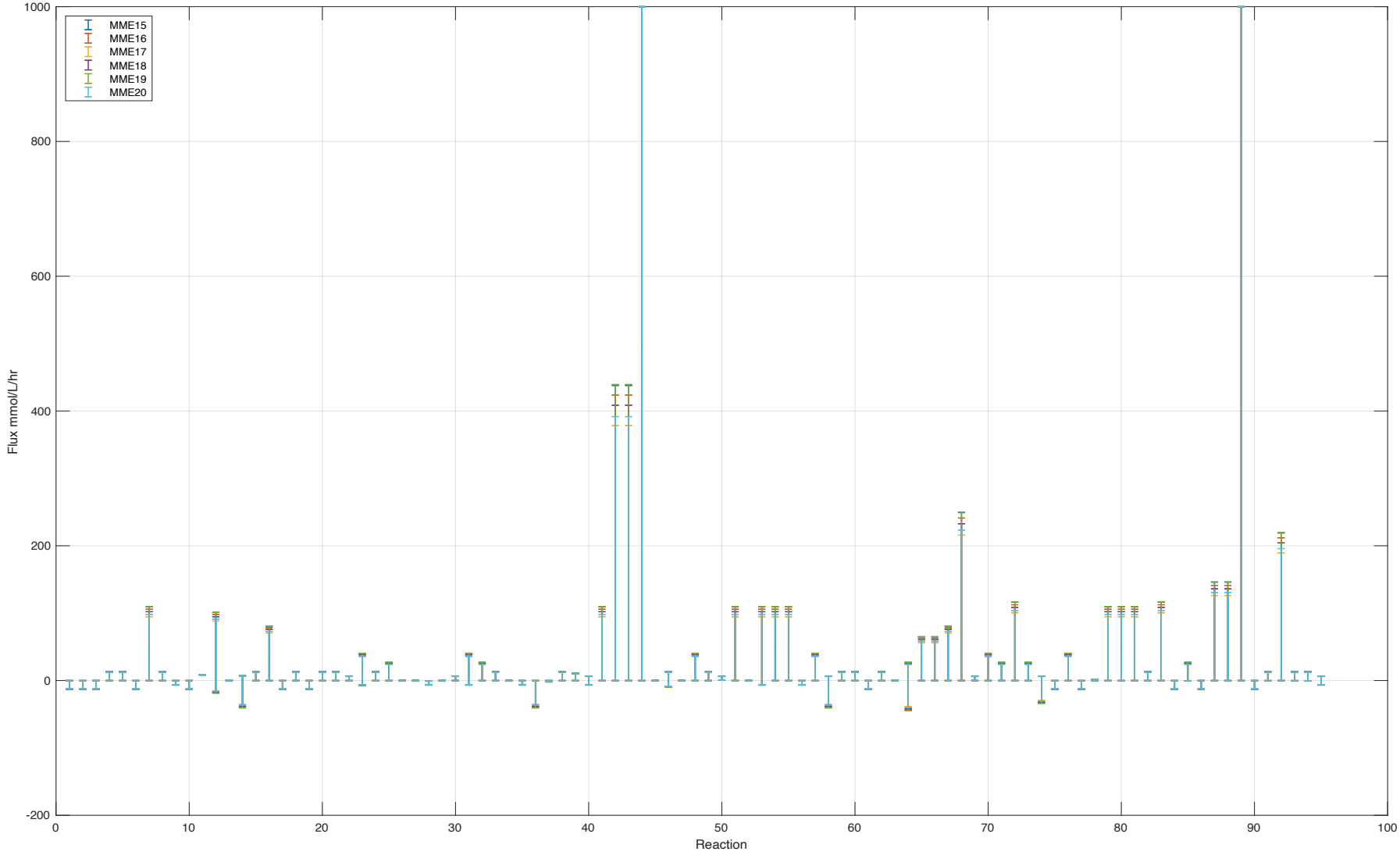

Figure 7.26 Flux range for each reaction in the batch phase generated by FVA

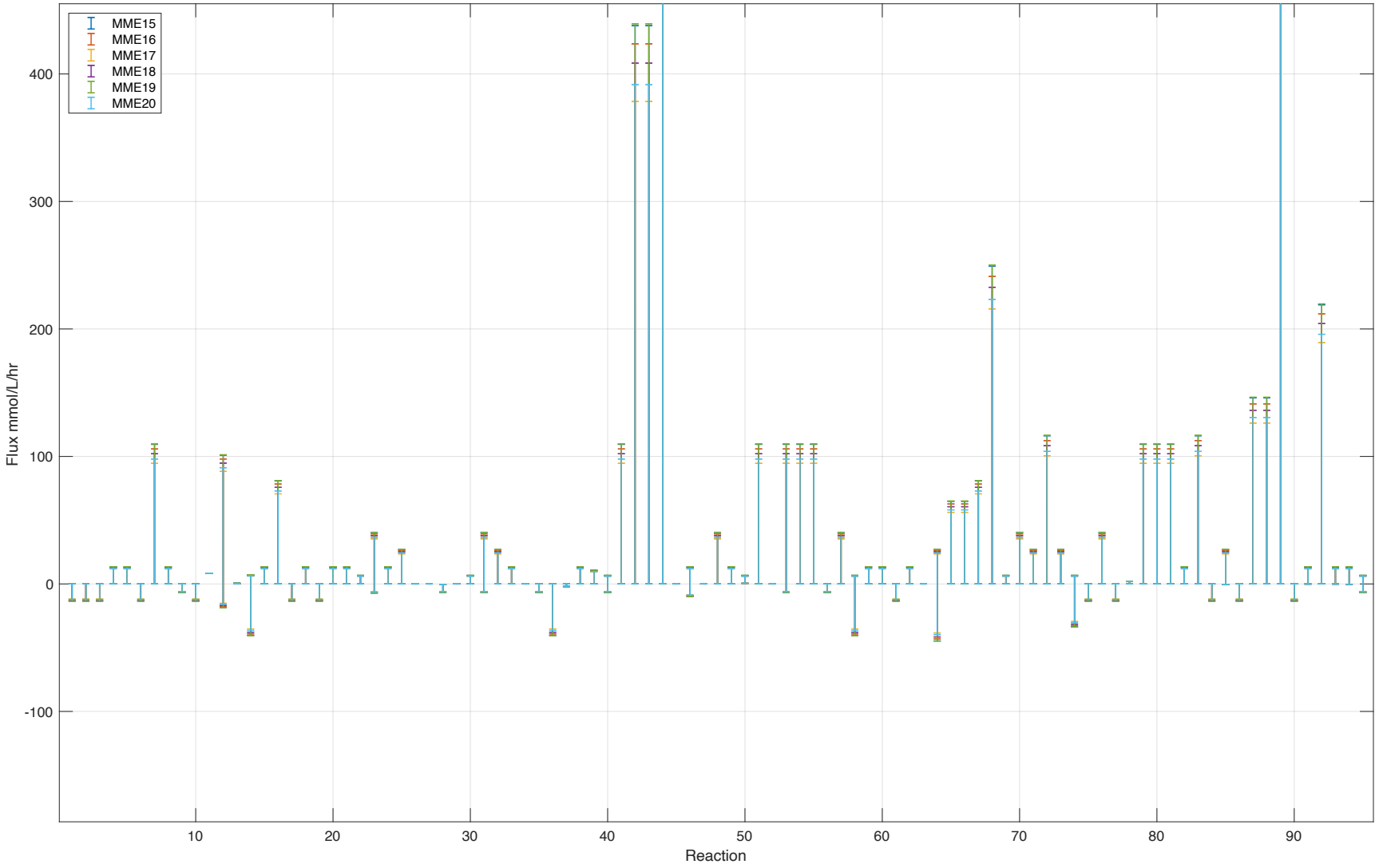

Figure 7.27 Flux range for each reaction in the batch phase generated by FVA with outliers removed from view

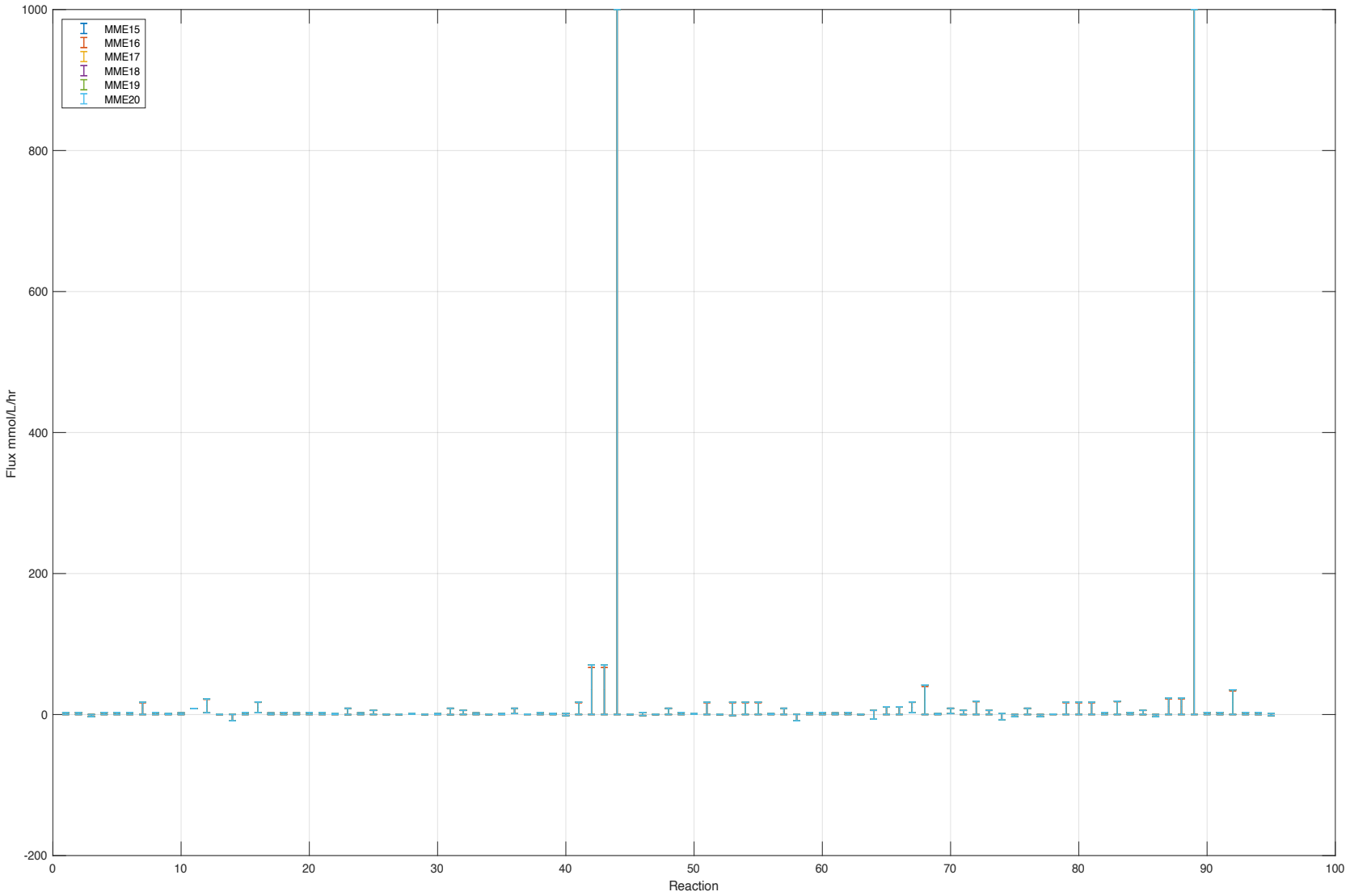

Figure 7.28 Flux range for each reaction in the fed batch phase generated by FVA

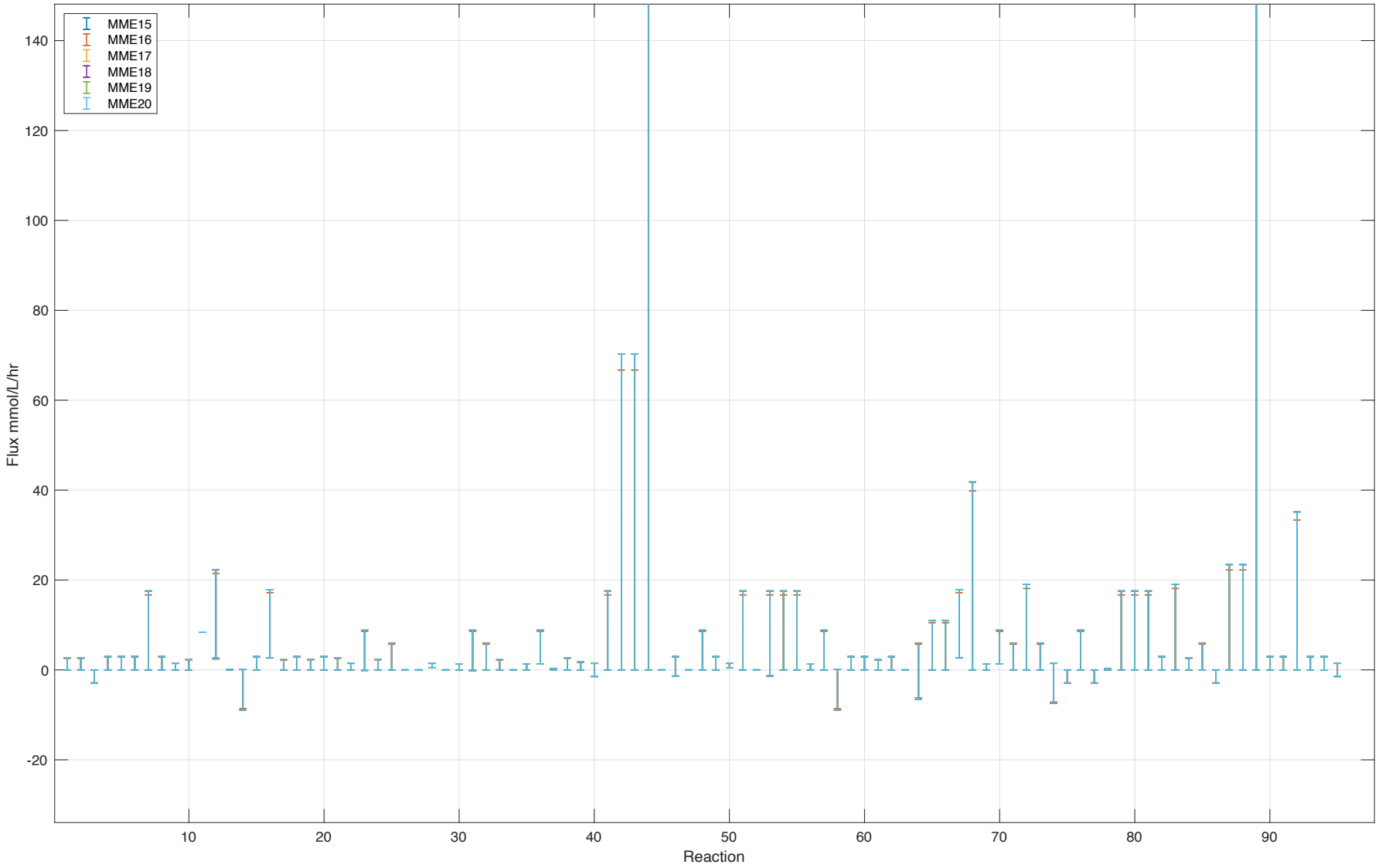

Figure 7.29 Flux range for each reaction in the fed batch phase generated by FVA with outliers removed from view

Efmtool finds 100274 EFMs for the batch system and 87163 for the fed batch. The batch phase therefore uses all EFMs possible in the *E. coli* core network [38]. The difference in these values is due to some reversible reactions operating in only one direction, which differs between the phases. The 13111 EFM reduction between the batch to the fed batch phase highlights that application of FVA in EFM enumeration in the MILP method would be beneficial. FVA provides the full range of unique solutions possible when maximising an objective. It, therefore, is more reflective of what could occur within a cell. As discussed in section 7.7, 7 reactions in both the batch and fed batch phase are unused to maximise biomass yield. These reactions can be removed from the MILP search space. In addition to this, some reversible reactions are only operating in one direction as the flux is 0 or negligible which will also reduce the search space.

The fed batch FVA results have been applied to the compressed *E. coli* core network used in the MILP method. In 10 minutes 896 EFMs were found with an efficiency of 0.9376. Compared to the compressed network results in Chapter 5, the number of EFMs found increased by 9.67% and efficiency improved by 4.32%. The number of EFMs found increased due to the reduced search space for MILP. Therefore, application of FVA to MILP in the future will aid in the enumeration of greater sets of EFMs; particularly those of which can occur within the experimental conditions. Elimination of reactions in the network allow for a fewer false EFMs to be found, which improves efficiency. Overall FVA is found to be a useful tool in improving MILP efficiency and solve time for EFM determination.

7.8 Conclusion

The *E. coli* cell data provided appears extensive at first glance but upon inspection lacks metabolite concentration data to allow for fully determinable flux analysis (metabolic flux analysis, MFA). The data was not originally collected for this project, hence the large amount of unmeasured data points. This chapter explored the biomass composition for all 6 fermentations, highlighting the differences in the three antigens: WT1, M72 and F4co. The limitations of the data only allowed flux balance and flux variability analysis to be performed. To use these techniques the specific uptake rate for glucose, the carbon source, and biomass growth rate were required. Both these rates were found for all fermentations in the batch and fed batch phases of production. The growth rate and uptake rate of glucose were lower during fed batch operations due to the cell entering the transition phase of life at ~40 hours.

FBA was found to reduce the network's reactions by nearly half. This however did not benefit the EFM enumeration as there were no EFMs in the batch phase and only 1 in the fed batch phase. This is due to the non-unique solution achieved in FBA, only providing the necessary reactions to maximise the biomass reaction. Therefore, FVA was used to find the range of flux for each reaction to maximise biomass yield. Only 7 reactions were found to be completely unused in the fermentations and some reactions were found to have negligible flux ($< \times 10^{-9}$). Application of these results allowed for a 9.67% increase in the number of EFMs found and an increase in efficiency of 4.32% for the MILP method presented within this thesis. The application of FVA creates a way to find more EFMs using MILP as it reduces the search space. Active EFMs for the *E. coli* cell could be determined if the data collected was more extensive as MFA could be used.

Chapter 8 Conclusion and Recommendations for Future Works

8.1 Overview of Aims

The aim for this thesis was to create a future-proof method of finding elementary flux modes (EFMs) so that genome scale models could be examined. EFMs found could then in turn be used to improve production efficiency and reduce waste in vaccine production by driving particular reactions. Each chapter had their own aim to contribute to the overall work.

Firstly, chapters 3 and 4 studied underdetermined and exactly determined flux analysis. Examination of flux analysis provided routes for reducing the search space for EFMs. Furthermore, a new integrated form of metabolic flux analysis was presented and the advantages of this method in reducing error discussed. Chapter 5's aim was to use a mixed integer linear programming (MILP) method to enumerate EFMs. Various networks were to be tested and the efficiency of the method compared to existing tools and MILP solvers. Then chapter 6 was aimed at reducing the solve time of the MILP method through techniques discussed in literature. Areas where MILP would be beneficial in the future were also to be highlighted in this chapter. The final chapter, chapter 7, was used to analyse the data provided by GlaxoSmithKline and assess the feasibility of determining flux and EFMs for the network. Application of flux analysis techniques were done, if possible, along with a critical analysis of where extra data was required to allow for active EFMs to be determined.

8.2 Result Overview

8.2.1 Chapter 3

This chapter studied flux balance (FBA) and flux variability analysis (FVA) for underdetermined systems. Through simulation it was proven that FBA results must be confirmed via experimental results to ensure their accuracy. However, key reactions to any metabolic network can be found via this optimisation problem. FBA results can also show which reactions are vital in the production of desirable and undesirable metabolites, creating a basic understanding of how environmental conditions should be set to drive reactions, e.g., reducing concentration of a substrate after a period. FVA produced a range of fluxes providing insight into how to maximise production of metabolites with low overall flux distribution. This

was found to be particularly helpful at highlighting how easily some metabolites, that negatively affect growth, are produced. FVA was found to be more useful than FBA as it did not just present one non-unique flux distribution. Therefore, this chapter highlighted that FVA could be used in conjunction with EFM enumeration to reduce the search space. Overall, these methods created good hypotheses for cell behaviour, but the lack of measured data created situations that were questionable and therefore required experimental confirmation and further data collection. A list of all possible measurable metabolites is needed, and this must be compared to those within the core network to allow for a more determinable system.

8.2.2 Chapter 4

This chapter used metabolic flux analysis (MFA) and integrated MFA (iMFA) to estimate flux and concentration change through an exactly determinable metabolic network. The main disadvantage to MFA proven in this chapter is its inaccuracy to real-time data. Through simulation and comparison with experimental data available, iMFA was found to mitigate this disadvantage. iMFA was proven to be a reliable technique suitable for industry to monitor transfer of material through all reactions in a metabolic network. If there was a reduction in material transfer, then it would show that reaction use was in decline. Knowledge of material transfer through reactions also enabled iMFA to be used in generating dynamic simulations of the cell's growth phase through the approximation of rate constants. iMFA offered an efficient method in approximating both the saturation constant and rate constant for a reaction rate in a Michaelis-Menten form. Whilst still offering a good prediction of substrate and product concentrations over time. To add to this, iMFA does not require behaviour assumptions of the cell to be made, apart from the pseudo-steady state assumption. Specific fluxes are not necessary in the prediction of intracellular material change, a major drawback of MFA.

8.2.3 Chapter 5

MILP was proven to be a viable method to enumerate EFMs and the work in this thesis improved on the solve time of MILP methods presented in literature via the addition of additional constraints. The commercial solver, efmtool, was quicker than MILP due to the many improvements made over the years to reduce memory usage and accelerate

computation time. However, this chapter highlighted areas of future improvements in solver efficiency and hardware which would reduce the MILP solve time.

8.2.4 Chapter 6

Areas of improvement mentioned in Chapter 5 were applied in Chapter 6. MILP caused memory storage clogging and reduced solve time for EFM enumeration, particularly at a large scale. Integer cuts were found to be unnecessary due to the extra constraints added to the MILP method in Chapter 5. Removal of these reduced the matrices size of the problem. This chapter also showed the introduction of sparse matrices throughout the code reduced the strain on the memory. Network compression was introduced and applied to various networks, including the *E. coli* core. Combining these techniques made it possible to solve over a 1000 EFMs in the *E. coli* core, a 302% increase on the method presented in Chapter 5. Flux data was also shown to be useful in reducing the search space. Zero fluxes indicated that reactions were not in operation, and this was used to prevent EFMs containing these reactions being found via MILP. Overall, this chapter emphasised the improvements needed to enable MILP to find more EFMs in a shorter period, whilst highlighting future methods that could be applied to apply the technique at genome scale.

8.2.5 Chapter 7

The data provided was for 6 fermentations with three different antigen productions: WT1, M72 and F4co. This chapter's key finding was the data provided lacked measurements allowing for it to be exactly determinable, restricting the flux analysis possible. As a result, only FBA and FVA could be performed. The specific uptake rate of glucose and the biomass growth rate were found for both batch and fed-batch operations. It was found that the growth rate and uptake rate of glucose were lower during fed batch operations due the cell entering the transition phase of life at ~40 hours. FBA reduced the network size but did not provide EFMs as the non-unique solution achieved in FBA only provided the necessary reactions to maximise the biomass reaction, which does not conform to the definition of an EFM. FVA was deemed a better flux analysis technique for the *E. coli* data and the flux range to maximise biomass was calculated. Only 7 reactions were found to be completely unused in the fermentations and some reactions were found to have negligible flux. Application of these

results allowed for a 9.67% increase in the number of EFMs found and an increase in efficiency of 4.32% for the MILP method presented within this thesis. The application of FVA created a way to find more EFMs using MILP as it reduced the search space.

8.3 Final Remarks and Future Scope

8.3.1 Chapter 3 and Chapter 4

These two chapters met aims set at the project's beginning; to understand and improve where possible on flux analysis techniques for determined and underdetermined systems. Flux analysis is well documented throughout literature, however, the errors from experimental data were found to skew results obtained in MFA. iMFA proved successful in reducing these errors whilst still providing meaningful results. Therefore, it is proposed to trail an integrated approach of FVA and FBA. An integrated form of FVA would provide a range of material transfers for every reaction required to maximise an output, e.g., biomass. Knowledge of the material transfer would allow reaction rates within the range to be predicted for each reaction. As the integrated FVA would find the minimum and maximum possible material transfer through a reaction, the change in the reaction rates for this minimum and maximum could also be analysed. This is of particular interest for the substrate saturation constant. If the specific growth rate is set to be half the maximum specific growth rate, the substrate saturation constant equals the concentration of the growth rate-limiting nutrient. Therefore, the substrate saturation constant is the concentration of growth rate-limiting nutrient that supports half the maximum specific growth rate [157]. So, if two enzymes are competing for one substrate, effectively two reactions, the route with the greater saturation constant will have the largest amount of relative flux. Collecting a range of these saturation constants will allow for evaluation of the most likely route travelled within the cell in minimum and maximum conditions.

8.3.2 Chapter 5 and Chapter 6

Chapter's 5 and 6 completed the main aim of the work to create a method to viably solve large scale EFMs. Currently the MILP is implemented into MATLAB as a standalone optimisation tool. Commercial optimisation packages, such as Gurobi and CPLEX, will offer an

improved solve time to MATLAB and should be considered a viable option in the future. These packages also offer a useful approach to Big M constraints which can cause instability within the optimisation. This work required flux data to estimate the Big M constraint for each reaction. However, the packages offer specially ordered sets (SOS) as an alternative to the Big M constraint. The Big M constraint is used to ensure that $\delta_i = 1, \Rightarrow v_i > 0$ and, $\delta_i = 0, \Longrightarrow v_i = 0$, equations (8.1) and (8.2). However, SOS do the same process without the additional of constraints, equation (8.3) [158]. SOS offer an always valid and numerically stable alternative to the Big M constraint. Using this method would reduce the number of constraints in the problem, speeding up solve time and reducing memory clogging.

$$v_i \leq M_i \delta_i \qquad (i = 1, \ldots., N_r) \tag{8.1}$$

$$\delta_i \leq v_i \qquad (i = 1, \ldots., N_r) \tag{8.2}$$

$$\delta_i, v_i \in SOS - 1 \tag{8.3}$$

Chapter 6 discussed the possibility of a parallelisation when solving for EFMs. In the future it would be beneficial to apply this method either on multiple cores on a computer or a high-performance computer, like the Rocket HPC at Newcastle University. The code would need to be reconfigured to ensure that no *"breaks"* existed as parallel computation is only possible within MATLAB if this is the case. Different processors could also be trailed within this study as MATLAB currently benchmark the Intel Core i9-12900 processor as being the most efficient at solving optimisation problems within their framework.

8.3.3 Chapter 7

Chapter 7 utilised the GSK data as set out in the aims to better understand the process and highlight measurements that are needed to take this work further. Unlike the concentrations of metabolites, the gene concentration is more extensive, therefore, further analysis of the gene concentration data could be performed. Gene changes over time were examined in this thesis but with each gene relating to a reaction it could in fact be used to reduce the search space for EFMs. The associated reactions with gene concentrations with little to no change over time can be assumed to be non-operational. This data could be used to reduce the EFM search space and ensure only EFMs that could be active are found.

To be able to perform more accurate FVA/FBA and to perform MFA, further metabolite concentrations are required. Tackling the *E. coli* core is an easier task experimentally than the genome due to less measurements being required. The *E. coli* core network consist of 72 metabolites, of which only four of these in the data set are measured. A total of 48 extracellular metabolites are measured and with the genome consisting of 1904 metabolites, emphasising that measuring data to fit the core is a much less intensive task which should be tackled first [20].

References

1. Provost A, Bastin G, Agathos S, Schneider Y-J. Metabolic design of macroscopic bioreaction models: Application to Chinese hamster ovary cells. Bioprocess and biosystems engineering. 2007;29:349-66.
2. Insights FB. Vaccines Market Size, Share, Growth. Pharmaceutical 2020.
3. Organisation WH. Global vaccine action plan 2011–2020 2011 [Available from: https://www.who.int/immunization/
global_vaccine_action_plan/GVAP_doc_2011_2020/en/.
4. Organisation WH. Child mortality and causes of death 2020 [Available from: https://www.who.int/gho/child_health/ mortality/mortality_under_five_text/en/.
5. Pollard AJ, Bijker EM. A guide to vaccinology: from basic principles to new developments. Nature Reviews Immunology. 2021;21(2):83-100.
6. Janeway CJT, P;Walport,M. The interaction of the antibody molecule with specific antigen. Immunobiology: The Immune System in Health and Disease 5th edition. New York: Garland Science; 2001.
7. Sallusto F, Lanzavecchia A, Araki K, Ahmed R. From vaccines to memory and back. Immunity. 2010;33(4):451-63.
8. Deng S, Liang H, Chen P, Li Y, Li Z, Fan S, et al. Viral Vector Vaccine Development and Application during the COVID-19 Pandemic. Microorganisms. 2022;10(7).
9. Rodrigues AF, Soares HR, Guerreiro MR, Alves PM, Coroadinha AS. Viral vaccines and their manufacturing cell substrates: New trends and designs in modern vaccinology. Biotechnol J. 2015;10(9):1329-44.
10. Maruggi G, Zhang C, Li J, Ulmer JB, Yu D. mRNA as a Transformative Technology for Vaccine Development to Control Infectious Diseases. Mol Ther. 2019;27(4):757-72.
11. Rosa SS, Prazeres DMF, Azevedo AM, Marques MPC. mRNA vaccines manufacturing: Challenges and bottlenecks. Vaccine. 2021;39(16):2190-200.
12. Burrell CJ, Howard CR, Murphy FA. Chapter 11 - Vaccines and Vaccination. In: Burrell CJ, Howard CR, Murphy FA, editors. Fenner and White's Medical Virology (Fifth Edition). London: Academic Press; 2017. p. 155-67.
13. Rappuoli R, editor The COVID Vaccine. GSK Postgraduate Research Conference 2022; Siena, Italy.
14. Rerks-Ngarm S, Pitisuttithum P, Nitayaphan S, Kaewkungwal J, Chiu J, Paris R, et al. Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. N Engl J Med. 2009;361(23):2209-20.
15. Peck M, Gacic-Dobo M, Diallo MS, Nedelec Y, Sodha SV, Wallace AS. Global Routine Vaccination Coverage, 2018. MMWR Morb Mortal Wkly Rep. 2019;68(42):937-42.
16. Organisation WH. Immunization Agenda 2030: A Global Strategy To Leave No One Behind. 2020.
17. Oliveira J, Reygaert WC. Gram-Negative Bacteria. StatPearls. Treasure Island (FL): StatPearls Publishing
Copyright © 2023, StatPearls Publishing LLC.; 2023.
18. Lim JY, Yoon J, Hovde CJ. A brief overview of Escherichia coli O157:H7 and its plasmid O157. J Microbiol Biotechnol. 2010;20(1):5-14.
19. Cole LA. Chapter 13 - Evolution of Chemical, Prokaryotic, and Eukaryotic Life. In: Cole LA, editor. Biology of Life: Academic Press; 2016. p. 93-9.

20. Kim H, Kim S, Yoon SH. Metabolic network reconstruction and phenome analysis of the industrial microbe, Escherichia coli BL21(DE3). PLOS ONE. 2018;13(9):e0204375.
21. Braun P, LaBaer J. High throughput protein production for functional proteomics. Trends Biotechnol. 2003;21(9):383-8.
22. Rosano GL, Ceccarelli EA. Recombinant protein expression in Escherichia coli: advances and challenges. Frontiers in Microbiology. 2014;5.
23. Nascimento IP, Leite LC. Recombinant vaccines and the development of new vaccine strategies. Braz J Med Biol Res. 2012;45(12):1102-11.
24. Sezonov G, Joseleau-Petit D, D'Ari R. Escherichia coli Physiology in Luria-Bertani Broth. Journal of Bacteriology. 2007;189(23):8746-9.
25. Zanghellini J, Ruckerbauer DE, Hanscho M, Jungreuthmayer C. Elementary flux modes in a nutshell: properties, calculation and applications. Biotechnol J. 2013;8(9):1009-16.
26. Schuster P. Taming combinatorial explosion. Proceedings of the National Academy of Sciences. 2000;97(14):7678-80.
27. Klamt S, Gagneur J, von Kamp A. Algorithmic approaches for computing elementary modes in large biochemical reaction networks. Syst Biol (Stevenage). 2005;152(4):249-55.
28. Hunt KA, Folsom JP, Taffs RL, Carlson RP. Complete enumeration of elementary flux modes through scalable demand-based subnetwork definition. Bioinformatics. 2014;30(11):1569-78.
29. Terzer M, Stelling J, editors. Accelerating the Computation of Elementary Modes Using Pattern Trees2006; Berlin, Heidelberg: Springer Berlin Heidelberg.
30. Terzer M. Large scale methods to enumerate extreme rays and elementary modes. 2009.
31. Terzer M, Stelling J. Large-scale computation of elementary flux modes with bit pattern trees. Bioinformatics. 2008;24(19):2229-35.
32. de Figueiredo LF, Podhorski A, Rubio A, Kaleta C, Beasley JE, Schuster S, et al. Computing the shortest elementary flux modes in genome-scale metabolic networks. Bioinformatics. 2009;25(23):3158-65.
33. van Klinken JB, Willems van Dijk K. FluxModeCalculator: an efficient tool for large-scale flux mode computation. Bioinformatics. 2015;32(8):1265-6.
34. von Stosch M, Rodrigues de Azevedo C, Luis M, Feyo de Azevedo S, Oliveira R. A principal components method constrained by elementary flux modes: analysis of flux data sets. BMC Bioinformatics. 2016;17(1):200.
35. Erdrich P, Steuer R, Klamt S. An algorithm for the reduction of genome-scale metabolic network models to meaningful core models. BMC Systems Biology. 2015;9(1):48.
36. Müller AC, Bockmayr A. Fast thermodynamically constrained flux variability analysis. Bioinformatics. 2013;29(7):903-9.
37. Gerstl MP, Jungeruthmayer C, Muller S, Zhanghellini J. Which sets of elementary flux modes form thermodynamically feasible flux distributions? FEBS. 2015.
38. Heirendt L, Arreckx S, Pfau T, Mendoza SN, Richelle A, Heinken A, et al. Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. Nature Protocols. 2019;14(3):639-702.
39. Horvat P, Koller M, Braunegg G. Recent advances in elementary flux modes and yield space analysis as useful tools in metabolic network studies. World Journal of Microbiology and Biotechnology. 2015;31(9):1315-28.

40. Conway M. Machine learning methods for detecting structure in metabolic flow networks. University of Cambridge, Computer Laboratory 2020.
41. Orth J, Fleming R, Palsson B. Reconstruction and Use of Microbial Metabolic Networks: the Core Escherichia coli Metabolic Model as an Educational Guide. EcoSal Plus. 2010.
42. Ben Yahia B, Malphettes L, Heinzle E. Macroscopic modeling of mammalian cell growth and metabolism. Appl Microbiol Biotechnol. 2015;99(17):7009-24.
43. Schuster S, Hilgetag C, Woods JH, Fell DA. Reaction routes in biochemical reaction systems: Algebraic properties, validated calculation procedure and example from nucleotide metabolism. Journal of Mathematical Biology. 2002;45(2):153-81.
44. Klamt S, Regensburger G, Gerstl MP, Jungreuthmayer C, Schuster S, Mahadevan R, et al. From elementary flux modes to elementary flux vectors: Metabolic pathway analysis with arbitrary linear flux constraints. PLoS Comput Biol. 2017;13(4):e1005409.
45. Poolman MG, Venkatesh KV, Pidcock MK, Fell DA. A method for the determination of flux in elementary modes, and its application to Lactobacillus rhamnosus. Biotechnology and Bioengineering. 2004;88(5):601-12.
46. Charnomordic B, David R, Dochain D, Hilgert N, Mouret JR, Sablayrolles JM, et al. Two modelling approaches of winemaking: first principle and metabolic engineering. Mathematical and Computer Modelling of Dynamical Systems. 2010;16(6):535-53.
47. Zamorano F, Vande Wouwer A, Jungers RM, Bastin G. Dynamic metabolic models of CHO cell cultures through minimal sets of elementary flux modes. Journal of Biotechnology. 2013;164(3):409-22.
48. Schwartz J-M, Kanehisa M. Quantitative elementary mode analysis of metabolic pathways: the example of yeast glycolysis. BMC Bioinformatics. 2006;7(1):186.
49. Wiback SJ, Mahadevan R, Palsson B. Reconstructing metabolic flux vectors from extreme pathways: defining the alpha-spectrum. J Theor Biol. 2003;224(3):313-24.
50. Kamp Av, Schuster S. Metatool 5.0: fast and flexible elementary modes analysis. Bioinformatics. 2006;22(15):1930-1.
51. Pfeiffer T, Sánchez-Valdenebro I, Nuño JC, Montero F, Schuster S. METATOOL: for studying metabolic networks. Bioinformatics. 1999;15(3):251-7.
52. Urbanczik R, Wagner C. An improved algorithm for stoichiometric network analysis: theory and applications. Bioinformatics. 2004;21(7):1203-10.
53. Trinh CT, Thompson RA. Elementary mode analysis: a useful metabolic pathway analysis tool for reprograming microbial metabolic pathways. Subcell Biochem. 2012;64:21-42.
54. Jevremović D, Trinh CT, Srienc F, Sosa CP, Boley D. Parallelization of Nullspace Algorithm for the computation of metabolic pathways. Parallel Computing. 2011;37(6):261-78.
55. Motzkin T, Raffia H, Thompson G, Thrall R. THE DOUBLE DESCRIPTION METHOD. 1953.
56. Fukada K, Prodon A. Double Description Method Revisted. Lecture Notes in Computer Science. 2005;1120.
57. Gagneur J, Klamt S. Computation of elementary modes: a unifying framework and the new binary approach. BMC Bioinformatics. 2004;5(1):175.
58. Wagner C. Nullspace Approach to Determine the Elementary Modes of Chemical Reaction Systems. Physical Chemistry. 2004:2425 - 31.

59. Ullah E, Yosafshahi M, Hassoun S. Towards scaling elementary flux mode computation. Briefinfs in Bioinformatics. 2019.
60. Schuster S, Hilgetag C. On elementary flux modes in biochemical reaction systems at steady state. Biological Systems. 1994;2(2):165-82.
61. Schuster S, Fell DA, Dandekar T. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. Nature Biotechnology. 2000;18(3):326-32.
62. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL. The large-scale organization of metabolic networks. Nature. 2000;407(6804):651-4.
63. Wagner A, Fell DA. The small world inside large metabolic networks. Proc Biol Sci. 2001;268(1478):1803-10.
64. Dusad V, Thiel D, Barahona M, Keun HC, Oyarzún DA. Opportunities at the Interface of Network Science and Metabolic Modeling. Frontiers in Bioengineering and Biotechnology. 2021;8.
65. Tun K, Dhar PK, Palumbo MC, Giuliani A. Metabolic pathways variability and sequence/networks comparisons. BMC Bioinformatics. 2006;7(1):24.
66. Larhlimi A, Blachon S, Selbig J, Nikoloski Z. Robustness of metabolic networks: a review of existing definitions. Biosystems. 2011;106(1):1-8.
67. Beguerisse-Diaz M, Bosque G, Oyarzun D, Pico J, Barahona M. Flux-dependent graphs for metabolic networks. NPJ Syst Biol Appl. 2018;4:32.
68. Asgari Y, Salehzadeh-Yazdi A, Schreiber F, Masoudi-Nejad A. Controllability in Cancer Metabolic Networks According to Drug Targets as Driver Nodes. PLOS ONE. 2013;8(11):e79397.
69. Mackie A, Keseler IM, Nolan L, Karp PD, Paulsen IT. Dead end metabolites--defining the known unknowns of the E. coli metabolic network. PLoS One. 2013;8(9):e75210.
70. Rohl A. Reduction, Projection, and
Simplification of Metabolic Networks.
71. Dantzig GB, editor Origins of the simplex method1990.
72. Pan P-Q. Simplex Method. 2014. p. 61-100.
73. Johnson DSG, M.R. Computers and Intractability. New York: W.H. Freeman and Company; 2002.
74. Schrijver A. Theory of linear and integer programming
. New York: Wiley; 1998.
75. Huang L, Chen X, Huo W, Wang J, Zhang F, Bai B, et al. Branch and Bound in Mixed Integer Linear Programming Problems: A Survey of Techniques and Trends2021.
76. Klamt S, Schuster S, Gilles ED. Calculability analysis in underdetermined metabolic networks illustrated by a model of the central metabolism in purple nonsulfur bacteria. Biotechnology and Bioengineering. 2002;77(7):734-51.
77. Palsson B, Orth J, Thiele I. What is flux balance analysis? Nature Biotechnology. 2010;28(3).
78. Llaneras F, Picó J. Stoichiometric modelling of cell metabolism. Journal of Bioscience and Bioengineering. 2008;105(1):1-11.
79. Provost A, Bastin G. Metabolic flux analysis: an approach for solving non-stationary underdetermined systems. MATHMOD; Vienna2006.
80. Rabinowitz JD, Enerbäck S. Lactate: the ugly duckling of energy metabolism. Nat Metab. 2020;2(7):566-71.

81. Fu T, Zhang C, Jing Y, Jiang C, Li Z, Wang S, et al. Regulation of cell growth and apoptosis through lactate dehydrogenase C over-expression in Chinese hamster ovary cells. Appl Microbiol Biotechnol. 2016;100(11):5007-16.
82. Wurm FM. Production of recombinant protein therapeutics in cultivated mammalian cells. Nat Biotechnol. 2004;22(11):1393-8.
83. Hansen HA, Emborg C. Influence of ammonium on growth, metabolism, and productivity of a continuous suspension Chinese hamster ovary cell culture. Biotechnol Prog. 1994;10(1):121-4.
84. Sun XM, Zhang YX. [Effects of ammonia on cell metabolism in the culture of recombinant CHO cells]. Sheng Wu Gong Cheng Xue Bao. 2001;17(3):304-9.
85. Traustason B. Amino Acid Requirements of the Chinese Hamster Ovary Cell Metabolism during Recombinant Protein Production. bioRxiv. 2019:796490.
86. Llaneras F, Picó J. Which metabolic pathways generate and characterize the flux space? A comparison among elementary modes, extreme pathways and minimal generators. J Biomed Biotechnol. 2010;2010:753904.
87. Bogaerts P, Vande Wouwer A. How to Tackle Underdeterminacy in Metabolic Flux Analysis? A Tutorial and Critical Review. Processes. 2021;9(9):1577.
88. Mahadevan R, Schilling CH. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metabolic Engineering. 2003;5(4):264-76.
89. Joshi C, Peebles C, Prasad A. Modeling and analysis of flux distribution and bioproduct formation in Synechocystis sp. PCC 6803 using a new genome-scale metabolic reconstruction. Algal Research. 2017;27:295 - 310.
90. Provost A, Bastin G. Dynamic metabolic modelling under the balanced growth condition. Journal of Process Control. 2004:717 - 28.
91. Szeliova S, editor Determination of CHO biomass composition. Constraint-based reconstruction and analysis (COBRA); 2018.
92. Yang S-T, Liu X, Zhang Y. Chapter 4 - Metabolic Engineering – Applications, Methods, and Challenges. In: Yang S-T, editor. Bioprocessing for Value-Added Products from Renewable Resources. Amsterdam: Elsevier; 2007. p. 73-118.
93. Portela R, Richelle A, Dumas P, Stosch Mv. Time Integrated Flux Analysis: Exploiting the concentration measurements directly for cost-effective metabolic network flux analysis. Microoraganisms. 2019;7.
94. Llaneras F, Pico J. A procedure for the estimation over time of metabolic fluxes in scenarios where measurements are uncertain and/or insufficient. BMC Bioinformatics. 2007;8:421.
95. Chassagnole C, Noisommit-Rizzi N, Schmid JW, Mauch K, Reuss M. Dynamic modeling of the central carbon metabolism of Escherichia coli. Biotechnol Bioeng. 2002;79(1):53-73.
96. Chen N, Bennett MH, Kontoravdi C. Analysis of Chinese hamster ovary cell metabolism through a combined computational and experimental approach. Cytotechnology. 2014;66(6):945-66.
97. Stephanopoulos GN, Aristidou AA, Nielsen J. Metabolic Engineering - Comprehensive Models for Cellular Reactions. In: Stephanopoulos GN, Aristidou AA, Nielsen J, editors. Metabolic Engineering. San Diego: Academic Press; 1998. p. 81 - 114.
98. Levenspiel O. Chemical Reaction Engineering. III ed1998. 704 p.
99. Brendel M, Bonvin D, Marquardt W. Incremental identification of kinetic models for homogeneous reaction systems. Chemical Engineering Science. 2006;61(16):5404-20.

100. Hellerstein MK, Neese RA. Mass isotopomer distribution analysis at eight years: theoretical, analytic, and experimental considerations. Am J Physiol. 1999;276(6):E1146-70.
101. Dai Z, Locasale JW. Understanding metabolism with flux analysis: From theory to application. Metabolic engineering. 2017;43(Pt B):94-102.
102. Chen K, Ying Z, Zhang H, Zhao L. Analysis of Least Absolute Deviation. Biometrika. 2008;95(1):107-22.
103. Sane R, Sinz M. Chapter 1 - Introduction of Drug Metabolism and Overview of Disease Effect on Drug Metabolism. In: Xie W, editor. Drug Metabolism in Diseases. Boston: Academic Press; 2017. p. 1-19.
104. Sun T, Kwok WC, Chua KJ, Lo T-M, Potter J, Yew WS, et al. Development of a Proline-Based Selection System for Reliable Genetic Engineering in Chinese Hamster Ovary Cells. ACS Synthetic Biology. 2020;9(7):1864-72.
105. Eppley RW, Rogers JN, McCarthy JJ. Half-saturation constants for uptake of nitrate and ammonium by marine phytoplankton 1. Limnology and oceanography. 1969;14(6):912-20.
106. Foutch GL, Johannes AH. Reactors in Process Engineering. In: Meyers RA, editor. Encyclopedia of Physical Science and Technology (Third Edition). New York: Academic Press; 2003. p. 23-43.
107. Schilling CH, Schuster S, Palsson BO, Heinrich R. Metabolic Pathway Analysis: Basic Concepts and Scientific Applications in the Post-genomic Era. Biotechnology Progress. 1999;15(3):296-303.
108. Palsson B, Famili I. The Convex Basis of the Left Null Space of the Stoichiometric Matrix Leads to the Definition of Metabolically Meaningful Pools. Biophysical Journal. 2003.
109. Bastin G. Lectures on Mathematical Modelling of Biological Systems2018.
110. Pey J, Planes FJ. Direct calculation of elementary flux modes satisfying several biological constraints in genome-scale metabolic networks. Bioinformatics. 2014;30(15):2197-203.
111. Chan SH, Ji P. Decomposing flux distributions into elementary flux modes in genome-scale metabolic networks. Bioinformatics. 2011;27(16):2256-62.
112. Kaleta C, de Figueiredo LF, Schuster S. Can the whole be less than the sum of its parts? Pathway analysis in genome-scale metabolic networks using elementary flux patterns. Genome Res. 2009;19(10):1872-83.
113. Röhl A, Bockmayr A. A mixed-integer linear programming approach to the reduction of genome-scale metabolic networks. BMC Bioinformatics. 2017;18(1):2.
114. Bixby RE, editor A Brief History of Linear and Mixed-Integer Programming Computation2012.
115. Nemhauser G, editor Intger programming: the global impact. EURO, INFORMS; 2013; Rome, Italy.
116. King A. Regression under a modern optimization lens. USA: Massachusetts Institute of Technology; 2015.
117. Tsu J, Díaz VHG, Willis MJ. Computational approaches to kinetic model selection. Computers & Chemical Engineering. 2019;121:618-32.
118. Guil F, Hidalgo JF, García JM. Boosting the extraction of elementary flux modes in genome-scale metabolic networks using the linear programming approach. Bioinformatics. 2020;36(14):4163-70.
119. Burgard AP, Nikolaev EV, Schilling CH, Maranas CD. Flux coupling analysis of genome-scale metabolic network reconstructions. Genome Res. 2004;14(2):301-12.

120. Hadicke O, Klamt S. Computing complex metabolic intervention strategies using constrained minimal cut sets. Metab Eng. 2011;13(2):204-13.
121. Tortajada M, Llaneras F, Picó J. Validation of a constraint-based model of Pichia pastoris metabolism under data scarcity. BMC Systems Biology. 2010;4(1):115.
122. Damiani C, Pescini D, Colombo R, Molinari S, Alberghina L, Vanoni M, et al. An ensemble evolutionary constraint-based approach to understand the emergence of metabolic phenotypes. Natural Computing. 2014;13(3):321-31.
123. Wang X, Xia K, Yang X, Tang C. Growth strategy of microbes on mixed carbon sources. Nature Communications. 2019;10(1):1279.
124. von Wulffen J, Sawodny O, Feuer R. Transition of an Anaerobic Escherichia coli Culture to Aerobiosis: Balancing mRNA and Protein Levels in a Demand-Directed Dynamic Flux Balance Analysis. PLoS One. 2016;11(7):e0158711.
125. Jurtshuk Jr P. Chapter 4 Bacterial Metabolism. In: Baron S, editor. Medical Microbiology. 4th ed. Texas1996.
126. Alteri CJ, Himpsl SD, Engstrom MD, Mobley HL. Anaerobic respiration using a complete oxidative TCA cycle drives multicellular swarming in Proteus mirabilis. mBio. 2012;3(6).
127. Arnold PK, Finley LWS. Regulation and function of the mammalian tricarboxylic acid cycle. J Biol Chem. 2023;299(2):102838.
128. Akram M. Citric acid cycle and role of its intermediates in metabolism. Cell Biochem Biophys. 2014;68(3):475-8.
129. Blackstock JC. CHAPTER 12 - The tricarboxylate cycle. In: Blackstock JC, editor. Guide to Biochemistry: Butterworth-Heinemann; 1989. p. 149-59.
130. Zhu MM, Goyal A, Rank DL, Gupta SK, Vanden Boom T, Lee SS. Effects of elevated pCO2 and osmolality on growth of CHO cells and production of antibody-fusion protein B1: a case study. Biotechnol Prog. 2005;21(1):70-7.
131. Optimisation G. Gurobi 9.5 Delivers Enterprise Features and Even Better Performance 2021 [
132. Edwards JS, Ramakrishna R, Palsson BO. Characterizing the metabolic phenotype: a phenotype phase plane analysis. Biotechnol Bioeng. 2002;77(1):27-36.
133. Burgard AP, Vaidyaraman S, Maranas CD. Minimal reaction sets for Escherichia coli metabolism under different growth requirements and uptake environments. Biotechnol Prog. 2001;17(5):791-7.
134. Jonnalagadda S, Srinivasan R. An efficient graph theory based method to identify every minimal reaction set in a metabolic network. BMC Systems Biology. 2014;8(1):28.
135. Vlassis N, Pacheco MP, Sauter T. Fast Reconstruction of Compact Context-Specific Metabolic Network Models. PLOS Computational Biology. 2014;10(1):e1003424.
136. Mojtaba T, Stephen PB. Metabolic network reductions. bioRxiv. 2019:499251.
137. Heinrich RS, S. The regulation of cellular systems. New York: Chapman & Hall; 1996.
138. Clement TJ, Baalhuis EB, Teusink B, Bruggeman FJ, Planqué R, de Groot DH. Unlocking Elementary Conversion Modes: ecmtool Unveils All Capabilities of Metabolic Networks. Patterns. 2021;2(1).
139. Stephanopoulos GN, Aristidou AA, Nielsen J. Metabolic Engineering - Regulation of Metabolic Pathways. In: Stephanopoulos GN, Aristidou AA, Nielsen J, editors. Metabolic Engineering. San Diego Academic Press; 1998.
140. Josephson B, Fraenkel D. Sugar Metabolism in Transketolase Mutants of Escherichia coli. Journal of Bacteriology. 1974.

141. Jung T, Mack M. Interaction of enzymes of the tricarboxylic acid cycle in Bacillus subtilis and Escherichia coli: a comparative study. FEMS Microbiology Letters. 2018;365(8).
142. Silhavy TJ, Kahne D, Walker S. The bacterial cell envelope. Cold Spring Harb Perspect Biol. 2010;2(5):a000414.
143. Xu Y, Zhao Z, Tong W, Ding Y, Liu B, Shi Y, et al. An acid-tolerance response system protecting exponentially growing Escherichia coli. Nature Communications. 2020;11(1):1496.
144. Thomas CM, Frost LS. Plasmid Genomes, Introduction to. In: Bell E, editor. Molecular Life Sciences: An Encyclopedic Reference. New York, NY: Springer New York; 2021. p. 1-20.
145. Gingold EB. Bacterial transformation. Methods Mol Biol. 1985;2:237-40.
146. Sugiyama H. WT1 (Wilms' tumor gene 1): biology and cancer immunotherapy. Jpn J Clin Oncol. 2010;40(5):377-87.
147. Oka Y, Tsuboi A, Elisseeva OA, Udaka K, Sugiyama H. WT1 as a novel target antigen for cancer immunotherapy. Curr Cancer Drug Targets. 2002;2(1):45-54.
148. Tait DR, Hatherill M, Van Der Meeren O, Ginsberg AM, Van Brakel E, Salaun B, et al. Final Analysis of a Trial of M72/AS01E Vaccine to Prevent Tuberculosis. New England Journal of Medicine. 2019;381(25):2429-39.
149. Efficacy and Safety of GSK Biologicals HIV Vaccine in Antiretroviral Therapy (ART)-naïve HIV-1 Infected Persons: Identifier NCT01218113 [Internet]. 2012 [cited 2023].
150. Srivastava AK, Gupta S. 2.38 - Fed-Batch Fermentation – Design Strategies. In: Moo-Young M, editor. Comprehensive Biotechnology (Second Edition). Burlington: Academic Press; 2011. p. 515-26.
151. Rodríguez-León JA, de Carvalho JC, Pandey A, Soccol CR, Rodríguez-Fernández DE. Chapter 4 - Kinetics of the Solid-State Fermentation Process. In: Pandey A, Larroche C, Soccol CR, editors. Current Developments in Biotechnology and Bioengineering: Elsevier; 2018. p. 57-82.
152. Deuster PA, Heled Y. Chapter 41 - Testing for Maximal Aerobic Power. In: Seidenberg PH, Beutler AI, editors. The Sports Medicine Resource Manual. Philadelphia: W.B. Saunders; 2008. p. 520-8.
153. Stephanopoulos GN, Aristidou AA, Nielsen J. CHAPTER 8 - Metabolic Flux Analysis. In: Stephanopoulos GN, Aristidou AA, Nielsen J, editors. Metabolic Engineering. San Diego: Academic Press; 1998. p. 309-51.
154. Zamorano F, Wouwer AV, Bastin G. A detailed metabolic flux analysis of an underdetermined network of CHO cells. Journal of Biotechnology. 2010;150(4):497-508.
155. Schaechter M. Escherichia coli and Salmonella 2000: the view from here. Microbiol Mol Biol Rev. 2001;65(1):119-30.
156. Ross AC, Chen Q, Ma Y. Chapter five - Vitamin A and Retinoic Acid in the Regulation of B-Cell Development and Antibody Production. In: Litwack G, editor. Vitamins & Hormones. 86: Academic Press; 2011. p. 103-26.
157. Owens JD, Legan JD. Determination of the Monod substrate saturation constant for microbial growth. FEMS Microbiology Reviews. 1987;3(4):419-32.
158. Snyder RD. Linear Programming with Special Ordered Sets. The Journal of the Operational Research Society. 1984;35(1):69-74.

Appendix A

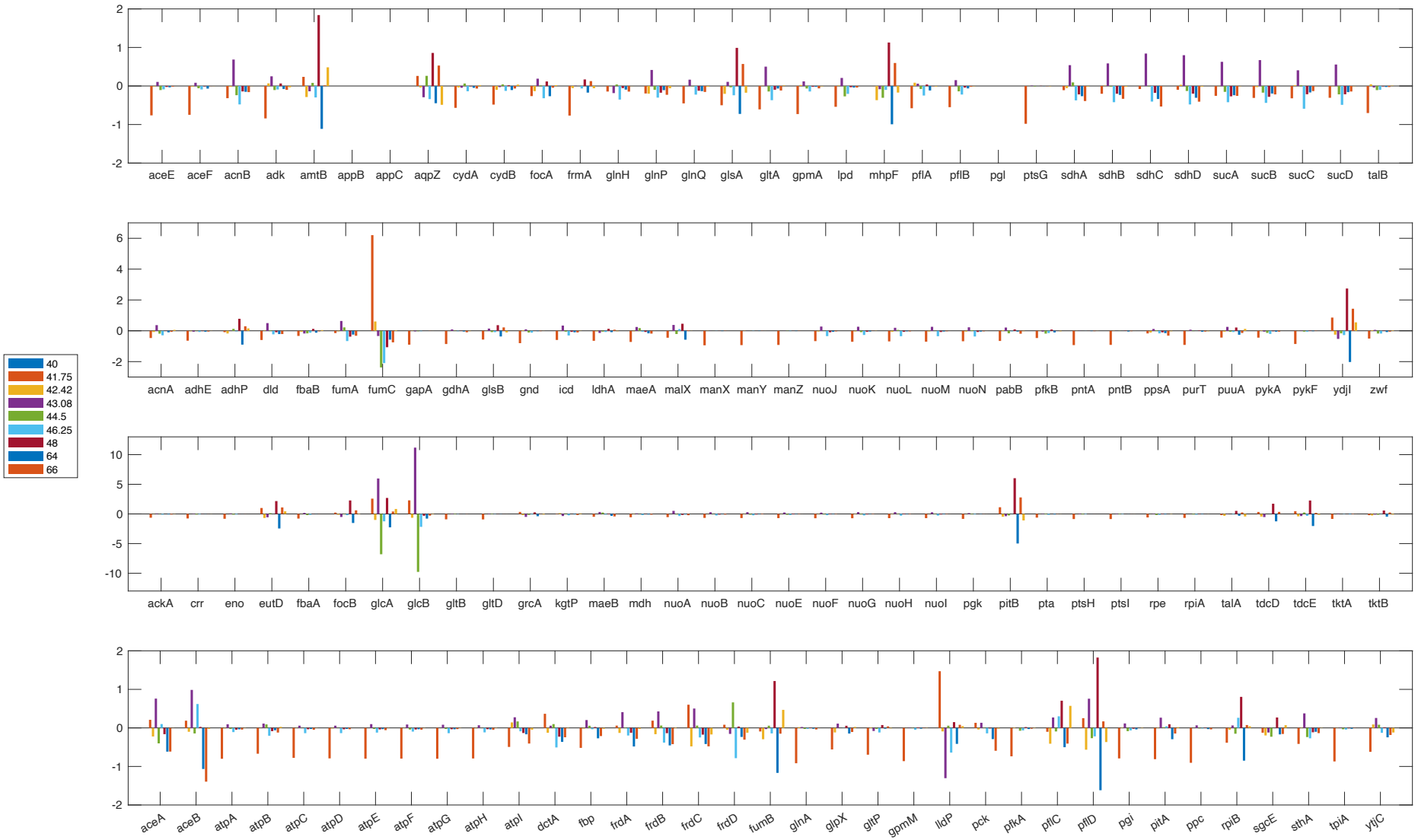

Figure A.1 MME15 relative gene expression at various sampling times

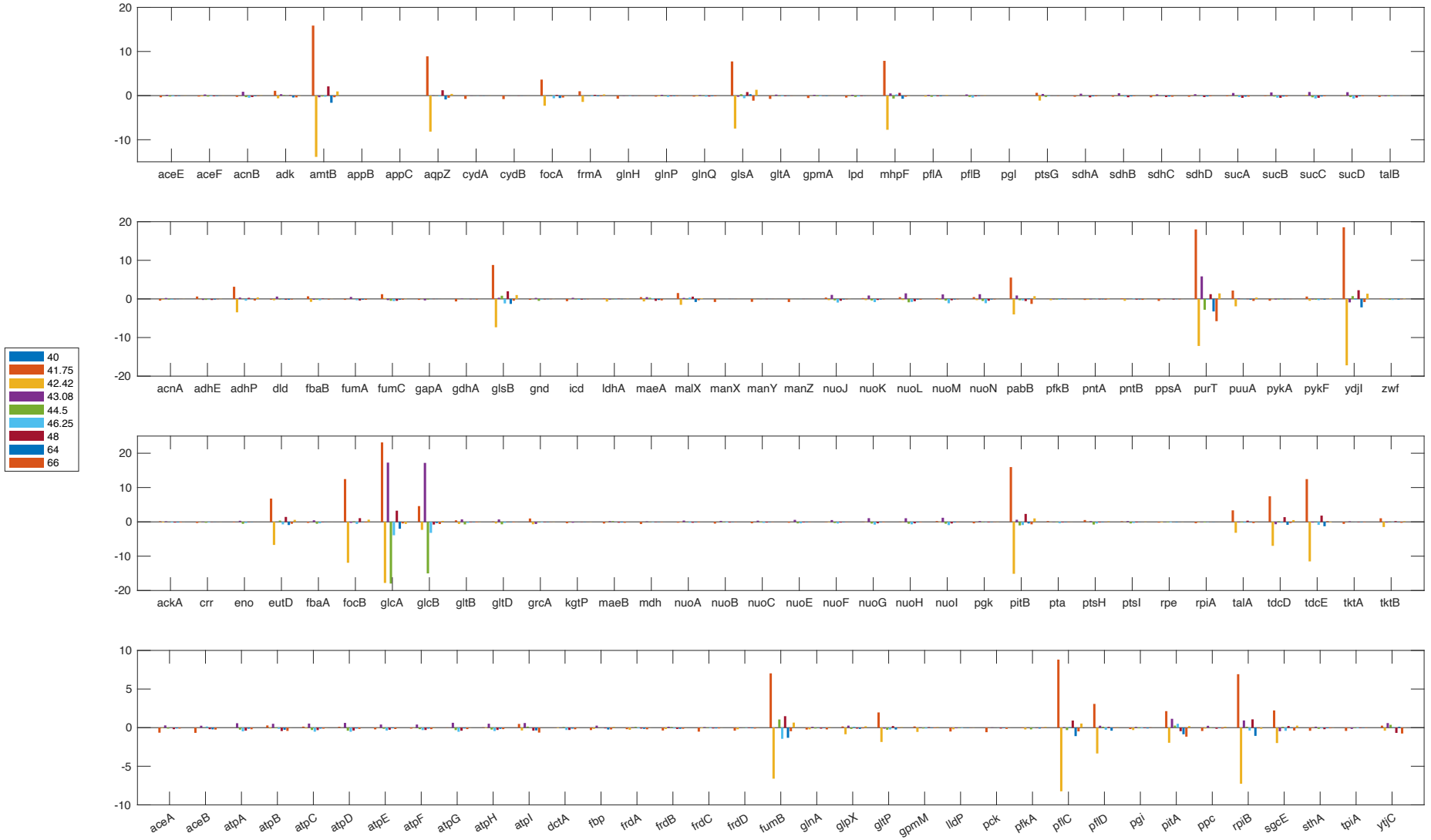

Figure A.2 MME16 relative gene expression at various sampling times

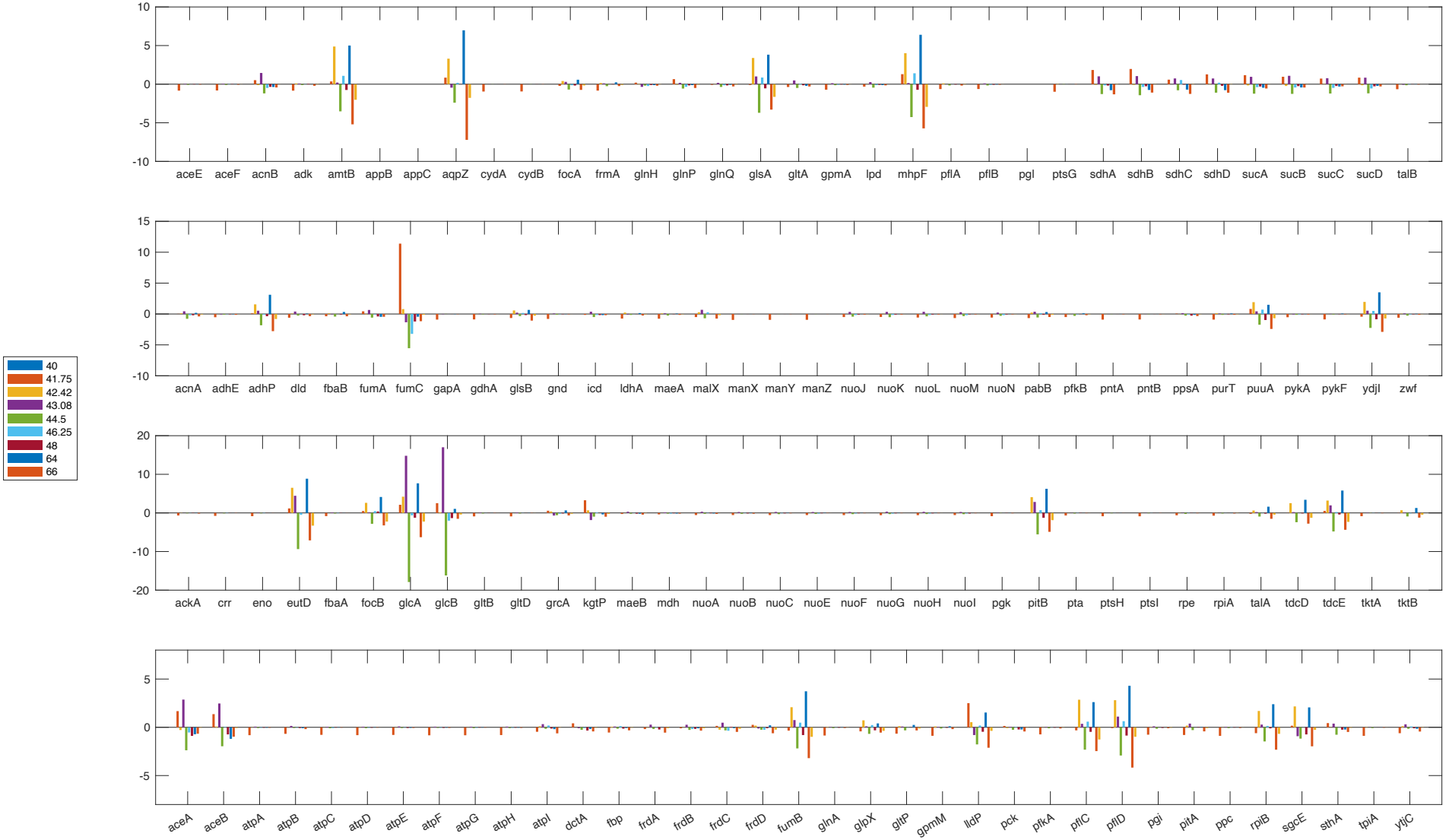

Figure A.3 MME17 relative gene expression at various sampling times

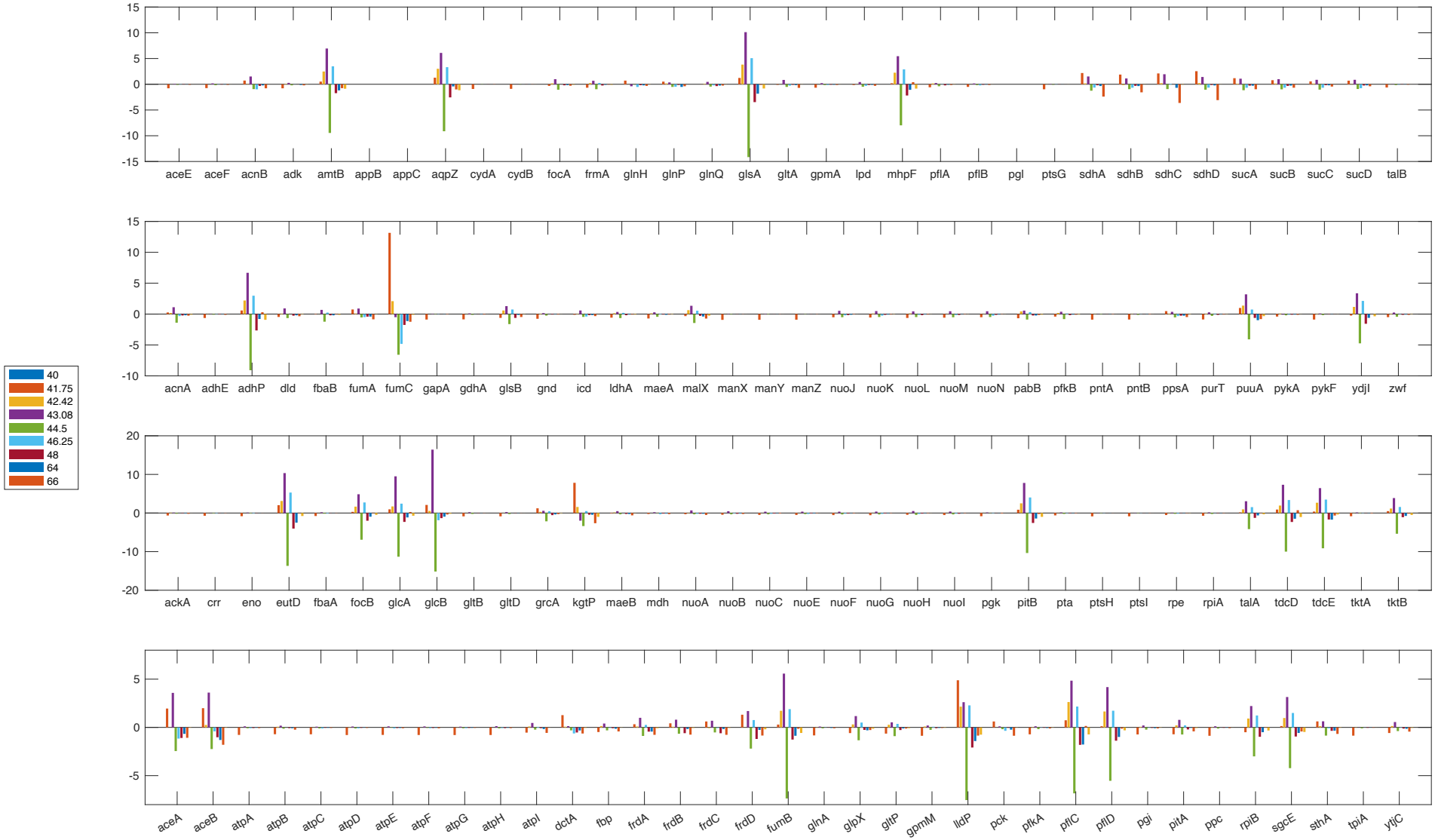

Figure A.4 MME18 relative gene expression at various sampling times