TÉCNICO LISBOA

Application of Minimal Cut Sets Algorithm to the Optimization of Plasmid and Recombinant Protein Production

Tiago Filipe dos Santos Roseiro

Thesis to obtain the Master of Science Degree in

Biological Engineering

Supervisor(s)

Prof. Isabel Cristina de Almeida Pereira da Rocha
Prof. Duarte Miguel De França Teixeira dos Prazeres

Examination Committee

Chairperson: Prof. Gabriel António Amaro Monteiro
Supervisor: Prof. Isabel Cristina de Almeida Pereira da Rocha
Member of the Committee: Dr. Sara Alexandra Gomes Correia

November 2018

Acknowledgments/Agradecimentos

First and foremost, I would like to thank my thesis supervisors Prof. Miguel Prazeres and Prof. Isabel Rocha for the time spent and valuable guidance and support. Their advice was extremely important throughout this project.

I want to take this opportunity to express my gratitude to Prof. Miguel Rocha and his whole lab for welcoming in their work and facilities during my stay in Braga. I want to give a special note to Vitor Vieira, for being extremely supportive and taking the time to teach me throughout this process. This research project would not have been finished successfully without his continuous support and readily available assistance (even when I was back in Lisbon). On another special note, I would like to thank Ana Sofia Ferreira for her last minute, but extremely valuable and crucial help, in this thesis. Without her, this project would not have the same quality.

I would like to gratefully acknowledge to all my friends that have been part of this 5-year journey. To my friends and colleagues Adriana and Cátia, I could not ask for better people to spend these 5 years working and laughing together. I wish you the best in your next new journeys. This one was truly amazing. To my Biodanish friends (Cat, Danylo, Gonçalo, Inês, Joana, Susana e Tomás) it was amazing and a pleasure to meet you all and, I hope we all meet again in a near future. I would like to give a special word to David and Cat, for their amazing support throughout this work. You guys were always there for me and I can not thank you enough for that.

Finalmente, gostaria de agradecer à minha familia, por tudo o que abdicaram e fizeram por mim para ser a pessoa que sou hoje. Obrigado a todos!

Abstract

Recombinant proteins (e.g. biopharmaceuticals, food processing enzymes, etc) are increasingly becoming more relevant in Biotechnology and Pharmaceutical industries. Current approaches for microbial strain optimization for industrial purposes rely heavily on modern Systems Biology. Regarding *in silico* methods, the most prominent currently comprise constraint-based modelling of cell metabolism, from central carbon metabolism to genome-scale models. Such approaches are used to solve metabolic engineering problems to fulfill an industrial objective and can be divided into phenotype prediction and pathway analysis (PA) methods. Unlike phenotype prediction, PA methods try to provide a more unbiased perspective but heavily depend on the complexity and scale of the model.

The aim of this work is to apply a PA method (minimal cut sets) to the optimization of plasmid and recombinant protein production. For this purpose, a novel implementation of an efficient algorithm for enumeration of minimal cut sets developed by Vieira (2015) was used.

The case study selected is based on a work performed by Pandey *et al.* (2018) and it involves interferon gamma production. Using an *E. coli* central metabolism model and a genome-scale model, different MCS enumeration problems were developed, for which knockout strategies were determined. An exploratory data analysis (principal component analysis and hierarchical clustering analysis) of the solutions was performed to select a few knockout sets for further analysis. The latter was performed to study the flux distributions and highlight different mechanisms of plasmid and/or product synthesis. From these analysis, it was possible to conclude that deletion of genes *pgi*, *pck* and *udhA/ptnAB* seem promising to increase *in vivo* plasmid and/or recombinant protein production. In addition, a further detailed analysis regarding genome-scale modelling would be beneficial to corroborate the results and add new knockout suggestions.

Keywords: Recombinant proteins; Constraint-based metabolic modelling; Flux balance analysis; Metabolic engineering; Pathway analysis; Minimal cut sets

Resumo

As proteínas recombinantes são cada vez mais relevantes nas indústrias biotecnológicas e farmacêuticas no que diz respeito, por exemplo, a agentes terapêuticos e à produção de enzimas importantes ao processamento alimentar.

As abordagens atuais para a otimização de estripes microbianas para fins industriais recorrem extensamente ao pensamento moderno da biologia de sistemas. Relativamente aos métodos *in silico*, estes usam, entre outros, modelos matemáticos baseados em restrições do metabolismo celular, desde metabolismo central a modelos de escala genómica. Tais abordagens são usadas para resolver problemas de engenharia metabólica para cumprir um objetivo industrial e podem ser divididas em métodos de previsão de fenótipos e de análise de vias metabólicas (AVM). Ao contrário da previsão de fenótipos, os métodos de AVM tentam fornecer uma perspetiva mais imparcial, mas são dependentes da complexidade e da escala dos modelos metabólicos.

Com este trabalho, o objetivo é aplicar um método AVM (*minimal cut sets - MCS*) na otimização da produção de plasmídeos e proteínas recombinantes. Para tal, utilizou-se uma nova implementação de um algoritmo eficiente para a enumeração de *minimal cut sets*, desenvolvida por Vieira (2015), num caso de estudo.

Este caso de estudo é baseado num trabalho realizado por Pandey *et al.* (2018) e envolve a produção de interferão gama. Usando um modelo de metabolismo central e um modelo à escala genómica de *E. coli*, foram desenvolvidos diferentes problemas de enumeração de *MCS*, para os quais foram determinadas estratégias de deleção. Uma análise exploratória de dados (análise de componentes principais e análise de clusters de métodos hierárquicos) das soluções foi realizada para selecionar alguns conjuntos de deleções para posterior análise. Esta última análise foi realizada com o intuito de estudar as distribuições de fluxo e destacar diferentes padrões e mecanismos de síntese de plasmídeos e/ou produtos.

A partir destas análises, foi possível concluir que deleção dos genes *pgi*, *pck* e *udhA/ptnAB* podem ser uma aposta promissora para aumentar a produção *in vivo* de plasmídeos e/ou proteínas recombinantes. Adicionalmente, uma análise mais detalhada utilizando modelos à escala genomica seria benéfica para corroborar os resultados encontrados e sugerir novos *knockouts*.

Palavras-Chave: Proteinas recombinantes; Modelação metabólica com base em restrições; *Flux balance analysis*; Engenharia metabólica; Análise de vias metabólicas; *Minimal cut sets*

Contents

List of Figures

List of Tables

List of Acronyms

ATPM Maintenance ATP

BPCY Biomass-product Coupled Yield

CBM Constraint-based Modelling

cMCS Constrained Minimal Cut Set

CMM Central Metabolism Model

COBRA Constraint-based Reconstruction and Analysis

DNA Deoxyribonucleic Acid

dNTP Deoxyribonucleotide Triphosphate

ED Entner-Doudoroff

EFM Elementary Flux Mode

ES Enzyme Subset

FBA Flux Balance Analysis

FVA Flux Variability Analysis

GSM Genome-scale Model

GUI Graphical User Interface

HCA Hierchichal Clustering Analysis

IDE Integrated Development Environment

IFG1 Insulin-like Growth Factor 1

IFNγ Interferon (gamma)

LP Linear Programming

MCS Minimal Cut Set

MFA Metabolic Flux Analysis

MW Molecular Weight

NGS Next-generation Sequencing

PA Pathway Analysis

PC Principal Component

PCA Principal Component Analysis

pFBA Parsimonious Enzyme Usage FBA

PPP Pentose Phosphate Pathway

PTS Phosphotransferase System

SBML Systems Biology Markup Language

TCA Tricarboxylic Acid

WT Wild-type

Chapter 1

Introduction

In this Chapter, some biological and computational background knowledge and general definitions regarding the topics addressed in this thesis are presented.

1.1 Recombinant Proteins

Recombinant proteins result from the expression of recombinant DNA that is introduced within a cell by genetic engineering methods. Over-expression of these therapeutically relevant proteins is increasingly a research area of interest for the Biotechnology and Pharmaceutical industry, as today over 100 recombinant proteins are used as therapeutic agents (Clark *et al.*, 2016). Most of these have human origin and some examples in clinical use are shown in Table 1.1. Additionally, many recombinant proteins are also industrially relevant, such as enzymes for laundry detergents and food processing.

Table 1.1: Examples of some therapeutically relevant proteins produced by recombinant DNA technology.

Protein	Function
Erythropoietin	Promotes red blood cells formation. Used to treat anaemia.
Factor VIII	Essential to blood-clotting. Used to treat haemophilia.
Insulin	Regulates carbohydrate metabolism. Used to treat diabetes.
Insulin-like growth factor 1 (IGF1)	Important role in child growth. Used to treat growth problems.
Interferon (beta)	Reduces multiple sclerosis relapse rates. Used to treat multiple sclerosis.
Interferon (gamma)	Important role in immunity. Used to treat chronic granulomatous disease.

The most commonly used host for over-expressing these proteins is *E. coli*, provided that post-translational modifications are not essential. This preference is based on *E. coli* genome being sequenced and extensively annotated; *E. coli* has a fast duplication time; cell culture is affordable; straightforward genetic manipulation strategies; and high potential to produce large protein amounts. In particular, *E. coli* B strains are more commonly used as expression hosts than *E. coli* K-12 derived strains. The latter is overlooked as it has propensity to accumulate acetate that may inhibit growth and the expression of heterologous proteins. However, B-type

strains have shown, for example, plasmid loss that completely arrests protein production and, as a result, non-B strains have gathered a lot of interest in recent years (Liu *et al.*, 2005; Waegeman *et al.*, 2015).

In addition to host-related problems, expressing these recombinant proteins at large-scale has its own obstacles such as, a high copy number of plasmids may lead to an increased metabolic burden, reducing host growth and often increasing plasmid instability (Liu *et al.*, 2015).

To improve the performance and yield of an expression host one can either manipulate its growth environment or alter its genetic architecture (using mutations and/or gene regulations). These changes may show effects on metabolic network flexibility, flux reaction efficiency and transcriptional regulation towards a desired product. One of the most common approaches tested by researchers to enhance and create an optimized microbial cell factory is to delete or add genes (Liu *et al.*, 2015; Pandey *et al.*, 2018).

With classical strain optimization methods, new microbial factories were developed based on the generation of mutants and selection of strains that have desirable phenotypic characteristics. These mutants were created by inducing random mutations through chemicals, radiation or transposons. Then, in a screening test, these mutants would grow in desired conditions and those that survived would be further optimized in new conditions or used for the purpose. However, in the start of the 21st century, with the development of systems biology and synthetic biology towards utilizing cellular network models combined with mathematical methods, metabolic engineering rationale had shifted. New computational methods, such as flux balance analysis (FBA) and constraint-based modelling (CBM), emerged and gave birth to a metabolic engineering era where strain optimization is first performed *in silico* and then tested *in vivo*. Instead of randomly screening numerous mutants, computational metabolic engineering is becoming increasingly a more direct and straightforward approach, that is continuously being improved throughout the years by the addition of new levels of complexity to the networks, as well as development of new methods and algorithms (Yang *et al.*, 2007).

1.2 Systems Biology

Systems biology is an interdisciplinary field that studies biological systems by describing the interactions within a system, instead of explaining individual mechanisms (Kirschner, 2005). In a more traditional perspective, biological studies follow a reductionist approach in which individual components of a living system are studied separately. This partitioning method requires a great workload amount of analysis and integration - specially with new generation technologies that present high throughputs - that could only be accomplished by innovative computational tools. Consequently, in the 21st century, there has been a shift towards an holistic and integrative methodology that has evolved not only from the reductionistic problem of dealing with bursting informations harnessed by high-throughput technologies, but also from lines of work that aim to study functional states of multiple components interactions simultaneously(Palsson, 2000; Westerhoff & Palsson, 2004).

Systems biologists utilize mathematical modelling methods to analyze biological interactions represented in different types of networks such as metabolic, transcriptional regulation

and signal transduction pathways to understand biological behaviour as a whole rather than compartmentalized. This field is becoming increasingly significant together with the improvement and development of high-throughput technologies and its objective is to enable the study of biological systems using the maximum amount of information possible at different levels of cell processes (Widlak, 2013). From cellular activities and metabolism to diagnosis and treatment of diseases, these are just some possible applications that can benefit from such approach (Raman & Chandra, 2009). In addition, other biotechnological fields such as genetic therapies and metabolic engineering can benefit immensely from systematic researches such as the one presented in this thesis. A schematic systems biology research cycle comprising main steps is depicted in Figure 1.1.

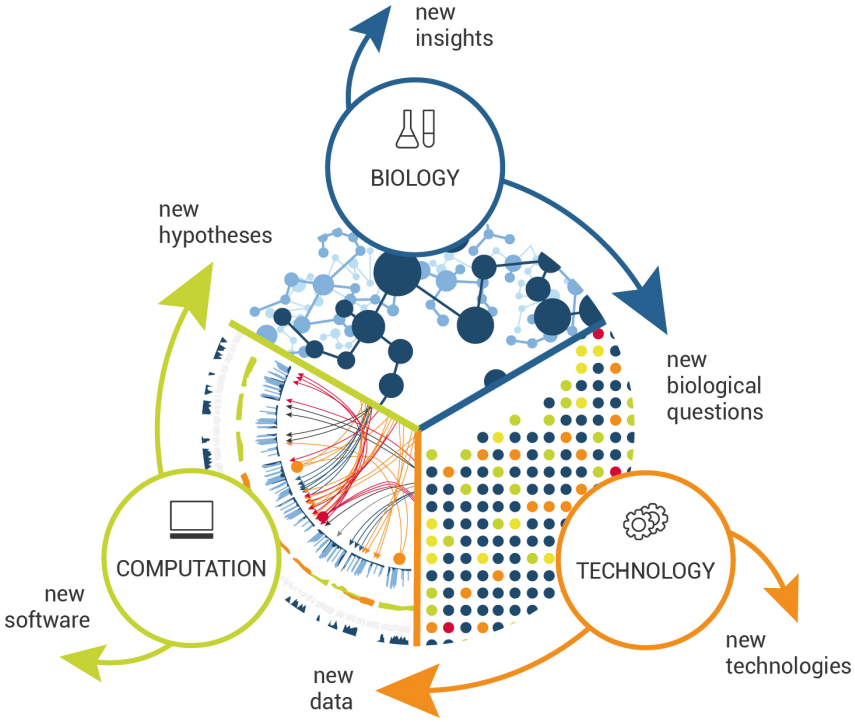

Figure 1.1: Systems biology research cycle. A new hypothesis is formulated and undergoes experimental design. From the lab experiments new data is generated and a model is constructed. From the latter the hypothesis is evaluated and reformulated and the process restarts until the model explains the data at maximum extent. From these cycles, new software and technologies may be developed. (From Institute of Systems Biology, 2018)

1.3 Metabolic Networks

Metabolic networks combine different levels of information in biological systems and describe relationships between metabolites and enzymes in a set of biochemical reactions (Castrillo *et al.*, 2013). Each reaction has key properties that are characterized as follows (Szallasi *et al.*, 2010):

- **Stoichiometry:** Specifies the molar ratios in which compounds participating in a reaction are consumed or produced. By convention, the stoichiometric coefficient of a compound is positive if it is produced when the reaction proceeds in its forward direction, and negative otherwise.

- **Reversibility:** In theory, all reactions are thermodynamically reversible. However, some can be considered irreversible due to their nearly unidirectionality. This information can be helpful in constructing accurate metabolic networks.

- **Enzymes:** Most biochemical reactions are characterized by the participation of an enzyme that facilitates or enables a reaction to proceed. Defining these enzymes allows correlating between network properties and features of the genome encoding those enzymes.

- **Kinetics:** Describes the dynamics based on the reaction mechanism and enzyme properties. These are defined by rate laws, which are mathematical expressions that describe reaction rates as a function of metabolite concentration and specific enzymatic kinetic parameters.

A metabolic network can be depicted as a graph where proteins/enzymes, that are edges of the network, interact with metabolites (nodes). An example of a metabolic network for central carbon flow in *E. coli* is given in Figure 1.2.

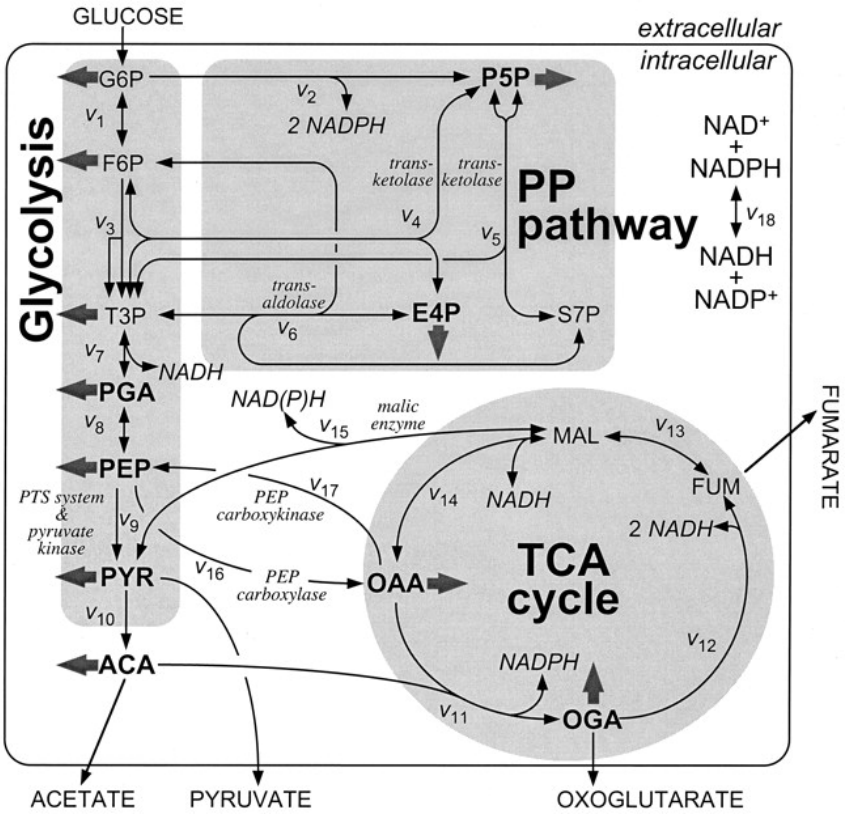

Figure 1.2: Typical bioreaction network of *E.coli* central carbon metabolism. Arrows indicate the assumed reaction reversibility. Fluxes to biomass building blocks are indicated by solid arrows (From Emmerling *et al.*, 2002).

The edges usually report an irreversible flux characterized by a uni-directional arrow (v_2 in Figure 1.2). Reversible reactions constitute two fluxes in opposite directions and are represented by a bi-directional arrow (v_1 in Figure 1.2). Intracellular reactions are edges that connect two groups of nodes (reactants and products), whereas exchange reactions only need one node with the edge connecting with the extracellular environment (Chalancon *et al.*, 2013).

Overall, metabolic networks play a critical role in numerous studies as this approach can provide the underlying reactions controlling all the physicochemical states of a cell in large scales.

1.3.1 Stoichiometric Matrix

The stoichiometric matrix, S, is essential to the mathematical representation of metabolic networks. It represents each metabolite as a row and each reaction as a column, where the numerical elements correspond to stoichiometric coefficients. This means that an (i, j) element represents the stoichiometric coefficients of metabolite i taking part in reaction j (Resendis-Antonio, 2013). Each entry depends on the role of metabolites in the reaction as follows:

$$S_{\text{ij}} = \begin{cases} a, & \text{number of molecules of } i \textbf{ produced} \text{ in reaction } j \\ -a, & \text{number of molecules of } i \textbf{ consumed} \text{ in reaction } j \\ 0, & \text{if metabolite } i \textbf{ does not take part} \text{ in reaction } j \end{cases}$$

Given a set of reactions, this matrix is constructed in a straightforward manner. As an example, Figure 1.3 shows the S matrix for a toy example with 10 metabolites and 8 reactions.

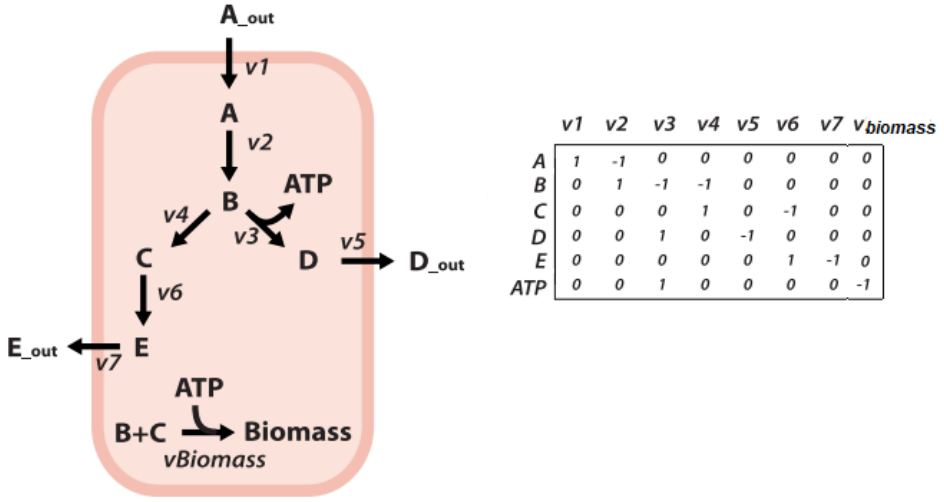

Figure 1.3: Toy example of a simplified metabolic network of a microorganism. The microorganism takes up metabolite A and produces Biomass, products D and E. On the right side, the corresponding stoichiometric matrix S, with rows corresponding to metabolites and columns to reactions (From Hanemaaijer *et al.*, 2015).

From this example, exchange reactions with the environment can be represented with internal metabolites that belong to the metabolic network (*A, B, C, D, E, ATP, Biomass*) and external metabolites that are considered pools or sources of internal metabolites (*A_out, D_out, E_out*).

If the same compound exists in multiple cellular compartments (for instance, ATP being present in cytosol and mitochondria in eukaryotes), it must be treated as a different metabolite for each compartment, meaning it must be given a separate row in the matrix (Becker *et al.*, 2007).

1.3.2 Genome-scale Metabolic Models

Metabolic models studies were initially restricted to small networks that only represented the central cell metabolism. One of the earliest studies to systematically analyze *E. coli* utilized a simplified constraint based model of acetate overflow (Majewski & Domach, 1990). Subsequent pre-genome-scale studies scaled-up to include reactions involved in central carbohydrate metabolism, amino acid and nucleotide synthesis to evaluate the biocatalyst production potential. As high-throughput sequencing methods became readily available, aligned with annotated content of *E. coli* in databases and detailed biochemical reviews, the information added to metabolic networks increased significantly (Baumler *et al.*, 2011). This expansion led to incorporation into a single systemic model of novel subsystems such as fatty acid synthesis, alternate carbon metabolism or cell wall synthesis, improving and promoting the metabolic reconstructions to the genome scale, ultimately leading to the reconstruction of genome-scale metabolic models (GSMs).

Genome-scale metabolic models have been reconstructed for over 150 organisms so far, including *E. coli* (Feist *et al.*, 2009). These reconstructions allow useful predictive calculations to be performed with high detail. GSMs of *E. coli* have existed for nearly twenty years as the first was released in 2000 by Palsson & Edwards, and this model continues to be expanded and updated today (McCloskey *et al.*, 2013). It accounts for the products of 660 metabolic genes, and has 627 reactions and 438 metabolites. It includes a biomass reaction based on the measured components of *E. coli* biomass that can be used to simulate growth (Edwards & Palsson, 2000). Table 1.2 highlights key events in the evolution of GSMs of *E. coli*.

Table 1.2: Evolution of GSMs of *E. coli* regarding date and version of model release. In addition, the number of model genes, metabolites and reactions is reported.

Date	Version	Model Genes	Metabolites	Reactions	Reference
11/05/2000	iJ660	660	438	627	Edwards & Palsson
04/09/2003	iJR904	904	625	931	Reed *et al.*
28/06/2007	iAF1260	1260	1039	2077	Feist *et al.*
07/01/2011	iCA1273	1273	1111	2477	Archer *et al.*
13/10/2011	iJO1366	1366	1136	2251	Orth *et al.*

GSMs can be applied to study, for instance, evolutionary processes, interspecies interactions and metabolic engineering problems (McCloskey *et al.*, 2013). Amidst these, metabolic engineering problems are some of the most studied using genome-scale models. This field tries to design new cells by using mathematical and experimental tools in metabolic analysis and modification. Thus, a systematic modelling can help shed some light into the complex nature of cellular metabolism and improve traditional methods for genetic engineering (e.g. random mutagenesis and screening for better phenotypes) by predicting cellular phenotypes from a systems level before *in vivo* implementation (Zhang & Hua, 2016; Yilmaz & Walhout, 2017). Figure 1.4 illustrates six fields and number of studies using *E. coli* metabolic GSMs until 2013. Since then, these numbers have seen an increase.

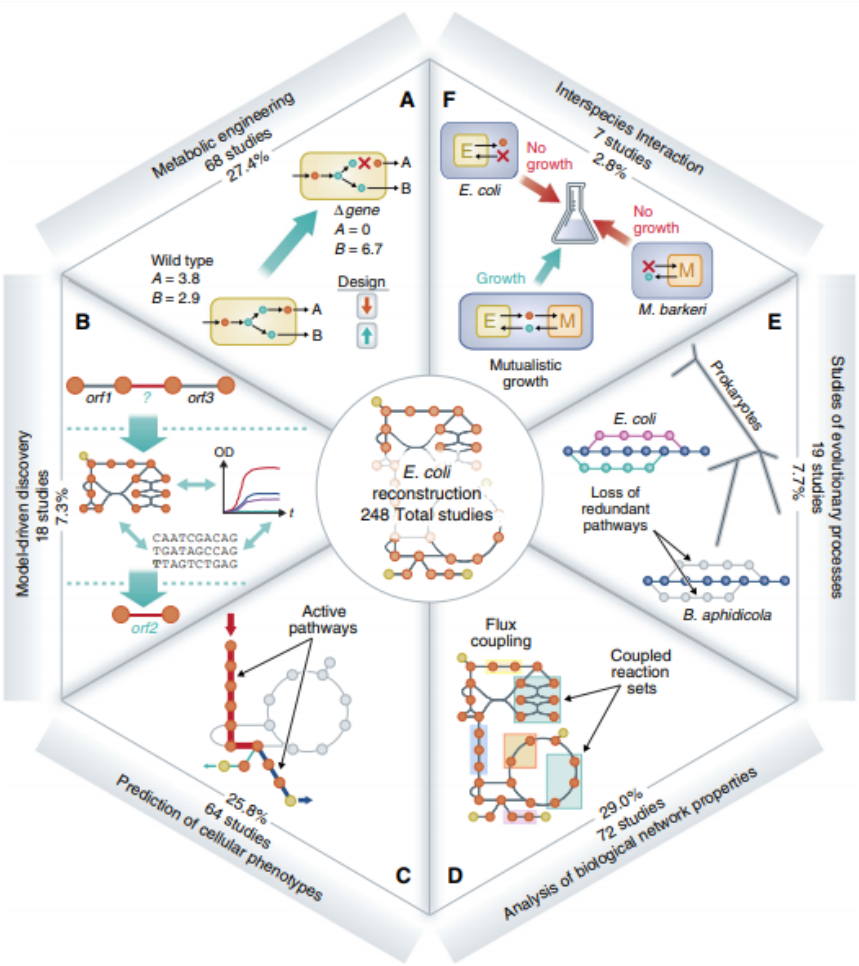

Figure 1.4: Six fields and number of studies for *E. coli* metabolic GSMs until 2013: **(A)** metabolic engineering, **(B)** model-driven discovery, **(C)** prediction of cellular phenotypes, **(D)** analysis of biological network properties, **(E)** studies of evolutionary processes and **(F)** interspecies interaction (From McCloskey *et al.*, 2013).

1.3.3 Mathematical Modelling Approaches

Combining the various levels of information from a biological system into a network enables the development of mathematical models that can be simulated under different conditions - dependant on the amount and type of data that is known and accessible.

Mathematical modelling is able to generate experimentally testable hypotheses on underlying mechanisms as well as predictions of cellular behaviour, thereby iteratively producing refined models and insight into the system (Kitano, 2002). It comprises several approaches to represent reality that ranges from global, yet coarse, views of cellular systems to detailed descriptions with a more limited scope. There are three main approaches to the mathematical modelling of cellular networks: (1) **interaction-based models** that are based on interactions alone; (2) **constraint-based models** that include constraints such as network topology, stoichiometry and reaction reversibilities; and (3) **mechanism-based models** where detailed reaction mechanisms and parameters are added (Stelling, 2004). A schematic summary of these mathematical modelling approaches is shown below in Figure 1.5.

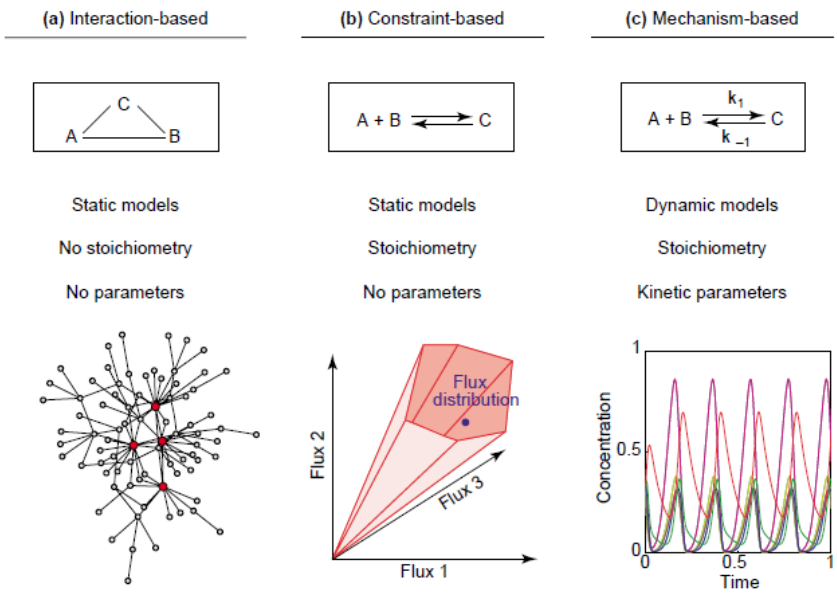

Figure 1.5: Cellular networks mathematical modelling approaches. In the top row, some key features of each method are presented. Schemes in the bottom row illustrate typical analysis results, namely **(a)** hubs (red circles) in a scale-free interaction network, **(b)** the cone of admissible flux distributions in a metabolic network constructed from the metabolic pathways (edges), and **(c)** dynamics in the concentrations of cellular components along time (From Stelling, 2004).

These methods will be discussed in the following subsections at a detailed level. Nevertheless, it is important to keep in mind that none of the approaches have the capability to cover the entire network complexity while maintaining a high level of detail and accuracy. However, mechanism-based modelling is the most obvious candidate for achieving a system-wide understanding; yet, it is not possible to scale to a complete genomic level (Figure 1.6).

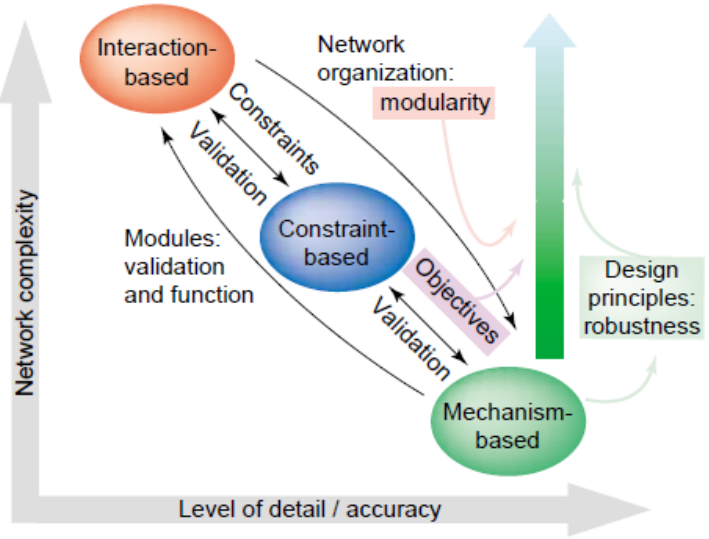

Figure 1.6: Mathematical modelling: scope and interactions. The three methods of modelling approaches are positioned according to the achievable degree of detail and accuracy, and the typical network sizes they can handle (network complexity). Black arrows refer to possible interactions. The green vertical arrow indicates the desirable progress towards genome-scale, mechanism-based models that allow for a system-level understanding from the genotype. Boxes and arrows in light colours visualize the contributions from all three approaches (From Stelling, 2004).

1.3.3.1 Interaction-based Models

Interaction-based models are static models that are based on interactions alone, not taking into account stoichiometry and kinetic parameters. These approaches highlight the existence of modules, that are semi-autonomous units performing distinct functions in cellular systems (Stelling, 2004). Network-level modules are defined variously as chemically isolated, operating on different time or spatial scales, robust, independently controlled, clusters in the graph-theory sense, and any or all combinations of the above (Hartell *et al.*, 1999).

Due to these models properties, such as its high coarseness and low detail level, this topological analysis appears particularly suited to reveal principles of cellular organization, but less able to handle network function and evolution (Wolf & Arkin, 2003).

1.3.3.2 Mechanism-based Models

Mechanism-based models are dynamic (or kinetic) models that attempt to describe cellular processes that are characterized by their dependence on time and susceptibility to external inputs on their states. In these models, mass balance equations that describe the temporal behaviour of all biochemical species are defined by using reaction kinetics and stoichiometry. For each metabolite involved in any reaction, one mass balance equation can be defined (Pfau *et al.*, 2011). In deterministic, continuous systems, these equations can be written as follows,

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = S \cdot v(x(t), u(t), \theta) \tag{1.1}$$

with their associated initial conditions,

$$x(0) = x_0(\theta) \tag{1.2}$$

where $x(t)$ denotes a vector of time-dependent metabolite concentrations (state variables), S a stoichiometric matrix and $v(x(t), u(t), \theta)$ a vector that is dependent on the state variables, a input vector $u(t)$ and a set of parameters θ. A system is then defined by a set of ordinary differential equations that are solved given a vector of initial conditions. The parameters appearing in the rate expression are also necessary to solve the equations and are often estimated using maximum likelihood, bayesian parameter estimates and by comparison with experimental data (Almquist *et al.*, 2014; Schaber *et al.*, 2009).

Although kinetic models excel at describing time-dependent cellular processes, the main challenge lies in producing high quality predictive models that can be used to improve cell performance. This is mainly due to incomplete and uncertain knowledge regarding kinetic rate expressions, as well as lack of experimental data to estimate valid parameters to characterize the complex metabolic network structure (Schaber *et al.*, 2009; Soh *et al.*, 2011).

1.3.3.3 Constraint-based Models

Constraint-based Models (CBMs) are static models where reactions stoichiometry and reversibility constraints are added to the network topology. This deals with the lack of kinetic information as these characteristics are largely available for metabolic networks (Szallasi *et al.*, 2010).

CBMs core idea is to incorporate physicochemical and biological constraints that limit the overall network behaviour and possible flux patterns confining the cellular phenotype to a set of feasible states (Orth *et al.*, 2010). Metabolism usually involves fast reactions and high turnover of substances. Therefore, it is assumed that, in longer time scales, metabolite concentration is stable meaning that the rates at which a metabolite is produced and/or consumed become constant over time. This generates an assumption that the system is time-invariant and in steady-state (Szallasi *et al.*, 2010). Applying this assumption to Equation 1.1 leads to:

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = 0 \tag{1.3}$$

and therefore,

$$S \cdot v = 0 \tag{1.4}$$

where v is no longer dependent on $u(t)$ and θ as it is in Equation 1.1 as these models do not account for kinetic rates. One trivial solution to this equation is $v = 0$ that represents thermodynamic equilibrium. However, one is looking for the remaining non-obvious solutions. Given a stoichiometric matrix S with $m \times n$ dimensions, usually there are more reactions n than the number m of internal metabolites. Consequently, this system will be undetermined ($m \leq n$) and all possible solutions are contained in a vector space called the null-space (or kernel) of S. Any point in this space can be described by a vector $v \in \mathcal{R}^n$ which is called a solution or flux distribution (Orth *et al.*, 2010).

Moreover, additional constraints can be added that are expressed by linear equations or inequalities. Regarding reaction capacity, one can define a range of acceptable flux values for each reaction. This is done by adding a upper bound ub_i and a lower bound lb_i to a reaction i, which will impose a maximum and minimum value, respectively.

$$lb_\mathrm{i} \leq v_\mathrm{i} \leq ub_\mathrm{i} \tag{1.5}$$

Capacities can also be translated in reaction reversibilities. If a reaction i is considered irreversible then $lb_\mathrm{i} \geq 0$, whereas if $lb_\mathrm{i} < 0$ the reaction is reversible. When there is no knowledge in regards to capacities the reaction rates limits are set to $\pm\infty$.

It is important to note that, if there is exact knowledge and measurements m_i of a flux

rate then $v_i = m_i$. This allows a reduction of degrees of freedom and, consequently reduces the solution space (Szallai *et al.*, 2010). An overall representation of CBMs is presented in Figure 1.7.

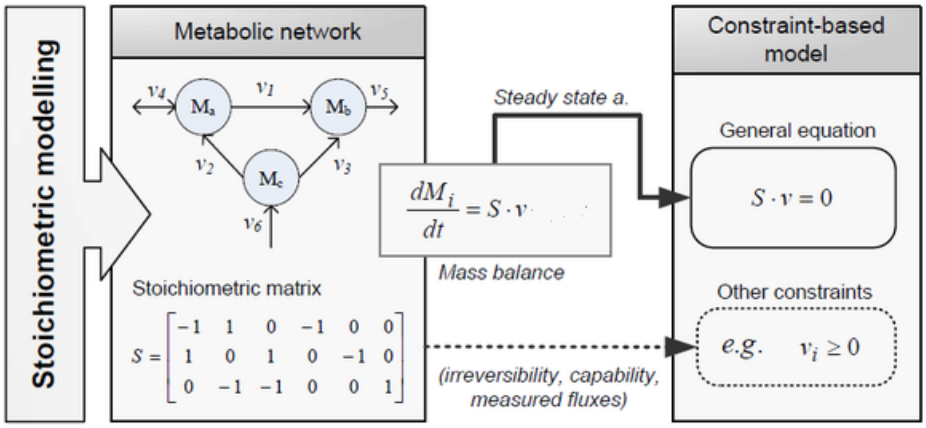

Figure 1.7: Principles of the stoichiometric modeling framework. Given a metabolic network, the mass balance around each intracellular metabolite can be mathematically represented with an ordinary differential equation. If we do not consider intracellular dynamics, the mass balances can be described by a homogeneous system of linear equations. Other constraints can be also incorporated to further restrict the space of feasible flux states of cells (Adapted from Estrada, 2010).

1.4 Phenotype Prediction

1.4.1 Flux Balance Analysis

Flux Balance Analysis (FBA) is a widely used phenotype prediction method to study biochemical networks. It calculates the flow of metabolites through a metabolic network, finding biologically relevant solutions whether by predicting the growth rate of an organism or the maximum production of a biotechonologically relevant product (Orth *et al.*, 2010).

To formulate a FBA problem, the first step is to mathematically represent the metabolic network. This is done by constructing a stoichiometric matrix that imposes constraints on the flow of metabolites through the network. Furthermore, additional constraints are added such as inequalities that impose boundaries in the system. Lastly, a linear objective function is required to solve the FBA problem. The latter function is defined by choosing a relevant biological objective in the study (Orth *et al.*, 2010). For example, in the case of growth prediction, the objective is biomass production, whereas in the case of product prediction, the objective is the reaction that produces it. Mathematically, an objective function is used to quantitatively define how much each reaction contributes to the phenotype and can be formulated as

$$Z = c^{\mathrm{T}} v \tag{1.6}$$

where c is the coefficient vector that defines the contributing weight of each flux in the objective function (Pfau *et al.*, 2011).

The metabolic network mathematical representation together with the objective define a system of linear equations, whose optimization problem can be generally solved using linear programming (LP) (Szallasi *et al.*, 2010). The general formulation for a simple FBA optimization problem is given as follows:

$$\begin{aligned} \max_{v} \quad & Z = f(v) \\ \text{s.t.} \quad & S \cdot v = 0 \\ & lb_{\mathrm{i}} \leq v_{\mathrm{i}} \leq ub_{\mathrm{i}} \end{aligned} \tag{1.7}$$

A summary representation of a FBA problem characterization is given below in Figure 1.8.

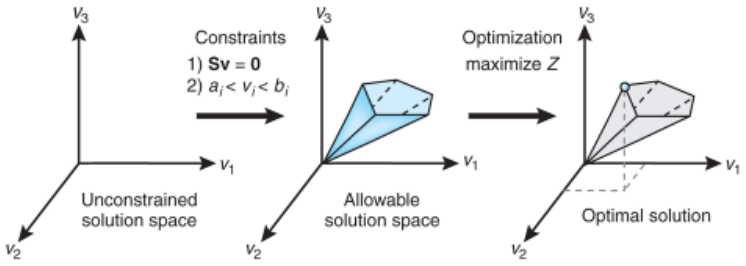

Figure 1.8: The conceptual basis of a FBA problem. With no constraints, the flux distributions may lie at any point in the solution space. When constraints are imposed by the stoichiometry matrix S and by the lower and upper bounds it defines an allowable solution space. Finally, through optimization of an objective function, a single optimal flux distribution can be determined that lies on the edge of the allowable solution space (From Orth *et al.*, 2010).

1.4.2 Flux Variability Analysis

The optimal solution to a FBA problem is rarely unique as there are other equally optimal existing solutions in the solution space. Flux variability analysis (FVA) is a derivative from FBA that aims to identify maximum and minimum fluxes through a reaction given an objective value, returning flux boundaries for each reaction.

Each flux is maximized and minimized and these objective values correspond to the true reaction limits in the metabolic network. Each reaction flux is computed using a double linear programming problem, meaning there is a maximization and a subsequent minimization, and these values correspond to the flux range in the metabolic network,

$$\begin{aligned} \min/\max_{v} \quad & v_{\mathrm{i}} \\ \text{s.t.} \quad & S \cdot v = 0 \\ & lb_{\mathrm{i}} \leq v_{\mathrm{i}} \leq ub_{\mathrm{i}} \end{aligned} \tag{1.8}$$

where v_{i} is the solution space for a reaction where $v_{\max}$ and $v_{\min}$ are calculated, containing the maximum and minimum feasible flux values, respectively (Gudmundsson & Thiele, 2010).

Reactions that present a low flux variability are more likely to be of a higher importance to the organism. Thus, FVA can be a promising technique for identifying important reactions and/or pathways in the model (Muller & Bockmayr, 2013).

1.4.3 Parsimonious Enzyme Usage FBA

Parsimonious enzyme usage FBA (pFBA) is a derivative from FBA where a second layer of optimization criteria is added making it a bilevel linear programming problem. It relies on the minimization of gene-associated protein cost while maintaining optimal growth. The pFBA optima represents set of genes associated with maximum growth as well as minimum-flux solutions, thereby predicting the most stoichiometrically efficient pathways.

This approach finds a flux distribution with minimum absolute values among the alternative optima, assuming that the cell attempts to achieve the selected objective function while allocating the minimum amount of resources (*i.e.* minimal enzyme usage).

1.5 Pathway Analysis

Pathway Analysis methods (PA), in contrast to methods such as FBA, are able to identify all metabolic flux vectors without imposing any objective function. Instead, they characterize the complete space of admissible steady-state flux distributions by functional/structural units alternately to searching specific flux vectors. Thus, PA attempts to provide an unbiased perspective of the theoretical limits of the network as a whole.

1.5.1 Nullspace Analysis

The nullspace is characterized by the kernel matrix K containing columns of linearly independent vectors that satisfy the condition given by Equation 1.4. Each column from this matrix is a basis vector that generates the complete solution space (Pfau *et al.*, 2011). From linear algebra rank-nullity theorem, it is possible to find the number of columns by determining the nullity of K using Equation 1.9.

$$nullity(S) = n - rank(S) \tag{1.9}$$

where n is the number of reactions in the system. Additionally, any flux distribution valid for Equation 1.4 can be constructed through linear combination of the columns from K,

$$r = K \cdot b \tag{1.10}$$

where b is a vector with the weight of each column in K.

Analysing the K matrix one can retrieve important information such as blocked reactions and enzyme subsets. Blocked reactions can be identified if their corresponding row i in K is a zero row. This is helpful since these reactions hardly have any function in the system and may be removed for practical reasons. Enzyme subsets (ES) (or coupled/correlated reaction set) are set of reactions that must operate together with a fixed reaction rate ratio. These can be identified from the null space matrix as the corresponding rows in K of a set of reactions from

the same ES can only differ by a scalar factor α,

$$v_{\mathrm{i}} = \alpha \cdot r, i = 1, ..., n \tag{1.11}$$

where v_{i} is the reaction rate vector for reaction i. These reactions are therefore linearly dependent (Szallasi *et al.*, 2010).

In nullspace analysis, thermodynamic constraints are not applied and thus, it is necessary to be careful when aiming for biologically relevant results as some proprieties, such as reaction reversibilities, are unconstrained and may be transgressed.

1.5.2 Convex Analysis

In convex analysis, contrarily to nullspace analysis, in addition to steady-state assumption, thermodynamic constraints are applied and the space of feasible flux distributions can be defined as follows,

$$P = \{v \in \mathbb{R}^n : S \cdot v = 0 \; ; \; I \cdot v \geq 0\} \tag{1.12}$$

where S is a $m \times n$ stoichiometric matrix, v a possible solution in the admissible space and I a diagonal $n \times n$ matrix with $I_{\mathrm{ii}} = 1$ if the flux i is irreversible (otherwise is 0).

This is a subset of the nullspace of S and in geometrical terms, this space of admissible flux distributions P, is a pointed convex polyhedral cone. This cone has a finite number of edges and is located in the positive orthant $\mathbb{R}^n_+$. By being convex, any vector within the cone (feasible solution) can be generated by non-negative linear combination of the vectors that generated the cone (which correspond to its edges) (Llaneras & Picó, 2010; Klamt *et al.*, 2017). A visual example of a convex polyhedral cone is given below in Figure 1.9.

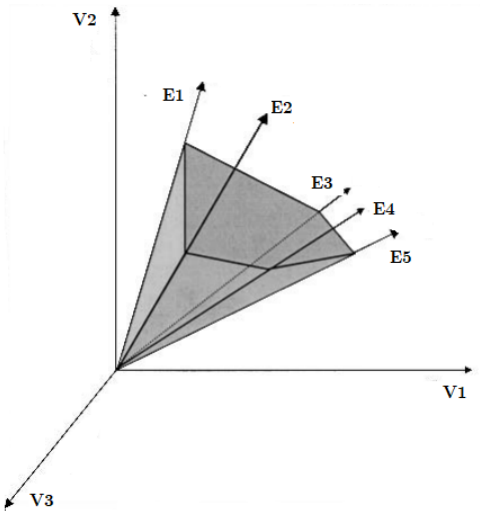

Figure 1.9: Representation of a pointed convex polyhedral cone for a metabolic network with three reactions (V1, V2 and V3). The admissible solution space (highlighted in grey) has positive or null reaction rates since it is strictly in the positive orthant $\mathbb{R}^3_+$. The edges (E1-E5) define the cone and can be used to describe any feasible flux distribution through linear combination. The cone basis represents the optimal solution space, obtained when using constraint-based approaches (Adapted from Papin *et al.*, 2002).

1.5.3 Elementary Flux Modes

Elementary Flux Modes (EFM) are flux distributions that are calculated by solving Equation 1.4 in conjunction with thermodynamic feasibility (Equation 1.12) and non-decomposability constraints. The support function $supp(v)$ provides a set of reaction indices from v with the condition that a reaction i can only be a part of $supp(v)$ if it has a nonzero flux value ($v_{\mathrm{i}} \neq 0$). Any elementary mode e is unique and minimal, in the sense that no reaction carrying a flux can be removed without violating the solving conditions. If a set of reactions constitutes an EMF then it fulfills the following proprieties (Schuster & Hilgetag, 1994; Schilling *et al.*, 2000; Schuster *et al.*, 2002):

- **Pseudo steady state:** According to Equation 1.4, no metabolite is consumed or produced in the overall stoichiometry. Hence, EFMs must belong to the nullspace of S.
- **Feasibility:** All fluxes have to be thermodynamically feasible and abide to their reaction reversibility. Hence, formally it requires that all rates $v_{\mathrm{i}} \geq 0$ if reaction $i \in$ **irrev**.
- **Non-decomposability:** This is the central property of EFMs and states that these flux distributions (or modes) represent the minimal functional units in a network. Hence, no reaction with a non-null flux value can be deleted from it, while still yielding a valid flux pattern. This feature is also known as genetic independence as this condition implies that the participating enzymes in one pathway are not a subset in another pathway.
- If e is an elementary flux mode, so is any $f = k \cdot e$ with $k > 0$.
- Every valid flux distribution v can be generated through linear combinations of support vectors that describe EFMs and/or scalars. These define the relative weight of each EFM in the flux distribution vector.
- Considering a set E of EFMs and a flux distribution vector v defined by Equation 1.13 where w is a vector with the relative weight of each EFM. If the EFMs in E are valid, $supp(E)$ can not contain reaction indices that are not already contained in $supp(v)$.

$$v = w \cdot E \; ; \; w \in \mathbb{R}_{0+}^{|E|} \tag{1.13}$$

In sum, each EFM can be defined as a unique, minimal set of reactions that support steady state operation of a metabolic network with irreversible reactions to proceed in appropriate directions (Trinh *et al.*, 2009). Thus, EFMs can be interpreted as the most elementary pathways of a metabolic system and are capable of providing concise information about the metabolic network, because they describe the possible (simplest) modes of operation of a system (Zanghellini *et al.*, 2013). The EFMs for a simple reaction network are shown in Figure 1.10.

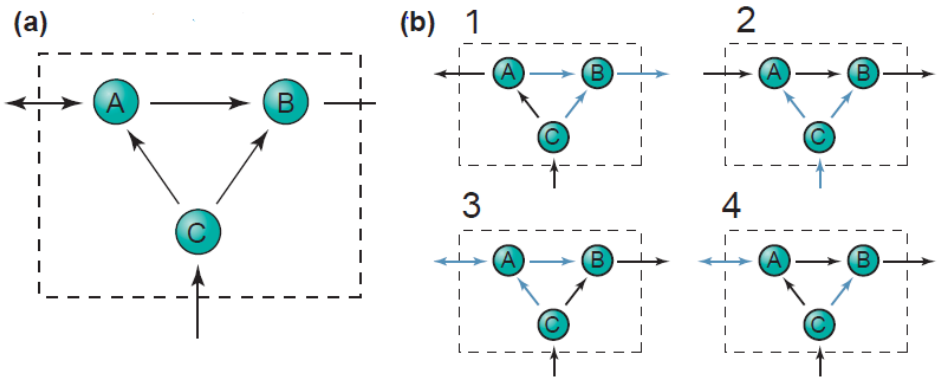

Figure 1.10: Simple example of a biochemical network and its elementary flux modes. The network consists of three metabolites (A, B and C), three internal reactions and three exchange reactions **(a)**, and there are four elementary modes **(b)**. The flux directionality is represented by black arrows, whereas reactions that do not have flux are represented by blue arrows. (Adapted from Papin *et al.*, 2004)

In the model, an EFM includes at least one input and one output that can be called net conversion. Identifying all the EFMs present in a model can be useful to identify which net conversion has the highest efficiency and which products are formed under each substrate. Similarly, EFMs performing undesired net conversions can also be identified (Pfau *et al.*, 2011; Zanghellini *et al.*, 2013). For instance, from the biochemical network from Figure 1.10, one can see that EFM2 and EFM4 have metabolite B as a product but what differs is the substrate, being A and C, respectively. This may be an indicator that this metabolic network is more robust when it comes to produce metabolite B.

Some of the most interesting applications of EFM analysis in Metabolic Engineering are: **(1) identifying** all range of possibles substrates and products, as well as finding ideal pathways to essentially modify and improve a desired metabolic capability (Szallasi *et al.*, 2010); **(2) establishing** the relative importance of a given reaction in a pathway. The higher number of EFMs that have the same reaction involved, the higher the likelihood of that reaction being a critical element to the metabolic system (Schuster & Hilgetag, 1994); and **(3) measuring** a pathway robustness through quantification. The number of EFMs that perform a given net conversion can be used as estimator to the pathway robustness (Szallasi *et al.*, 2010).

1.5.4 Minimal Cut Sets

Minimal Cut Sets (MCS) are a complementary concept to EFMs. A cut set is a set of reactions that need to be removed to inactivate a specified target reaction or, in another perspective, reactions whose deletions leads to network failure (considering a target reaction). A cut set becomes a MCS and is minimal in the sense that removing any subset of it from the network is not sufficient to maintain the target reaction inactivation. This means that by removing one reaction from the MCS prevents it from being a cut set anymore (Klamt & Gilles, 2004; Clark & Verwoerd, 2012).

To illustrate the MCS concept, consider the example network shown in Figure 1.11.

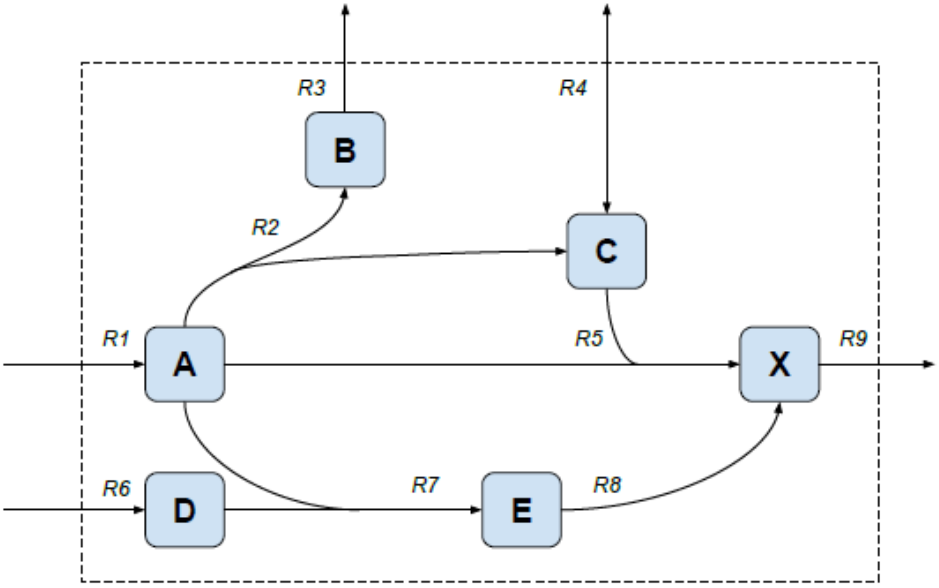

Figure 1.11: Biochemical network example. The network consists of six metabolites (A, B, C, D, E and X) and nine reactions (R1, R2, R3, R4, R5, R6, R7, R8 and R9) (Adapted from Klamt & Gilles *et al.*, 2004).

Assuming that one wants to block the production of metabolite X, a trivial solution is to cut the target reaction (*R9*) itself. However, it is not a biological reasonable strategy as *R9* is an exchange flux (pseudo-reaction) and, therefore, might not have the corresponding genes to be candidates for deletion. To find all MCSs that block *R9*, the minimal set of reactions that disable all EFMs must be found (Klamt & Gilles, 2004; Klamt, 2006). The set of EFMs and MCSs that block *R9* in the network depicted in Figure 1.11 are in Figure 1.12.

		R1	R2	R3	R4	R5	R6	R7	R8	R9
Elementary modes	EM1	1	1	1	-1	0	0	0	0	0
	EM2	1	0	0	0	0	1	1	1	1
	EM3	2	1	1	0	1	0	0	0	1
	EM4	1	0	0	1	1	0	0	0	1
Minimal cut sets	MCS1									•
	MCS2	•								
	MCS3					•	•			
	MCS4					•		•		
	MCS5					•			•	
	MCS6		•		•		•			
	MCS7			•	•		•			
	MCS8		•		•			•		
	MCS9			•	•			•		
	MCS10		•		•				•	
	MCS11			•	•				•	

Figure 1.12: Elementary modes and minimal cut sets that block *R9* from the network in Figure 1.11. Elementary flux modes that carry flux through *R9* are highlighted in grey. (Adapted from Klamt & Gilles *et al.*, 2004).

Another example is MCS5, where *R5* and *R8* are deleted. This cut set is sufficient enough to prevent production of X. Moreover, removing only *R5* or only *R8* will allow the flux to go through *R9* again. Thus, this cut is a MCS because no subset of {*R5,R8*} would be a subset anymore. If there was a new subset, one would have a sub-optimal cut set instead of a minimal cut set. Additionally, MCS2 is the only cut set with one reaction (apart from the trivial solution). This could be an excellent suitable candidate since it seems to be essential to synthesize metabolite X (Klamt & Gilles, 2004; Klamt, 2006).

In a Metabolic Engineering context, MCSs are useful to predict sets of genes which should be knocked out in order to inactivate a particular metabolic reaction based on the smallest set of reactions to achieve this goal. Alternatively, in a scenario where a given metabolite is desired to be produced it is feasible to calculate the MCS using additional constraints (Klamt & Gilles, 2004; Clark & Verwoerd, 2012).

A limitation to using MCSs is that they might disable not only undesired reactions but also desired functions. For instance, one MCS may block the synthesis of an undesired product, while at the same time removing the substrate uptake for a reaction where a metabolite of interest is being produced. To account for the need of keeping some reactions/EFMs intact, the concept of *constrainted* MCS (cMCS) can be introduced. Formally, a set of desired EFMs , D, is defined alongside a set of undesired modes (target), T. An admissible MCS is reached when all target modes T are hit, while preserving a minimum number n of desired EFMs. This results in a set of reactions ready to be deleted from the network and that are still guaranteed to provide the desired functionalities (Hadicke & Klamt, 2011).

1.6 Motivation and Objectives

This thesis has its starting grounds on a previous study done by Pandey *et al.* (2018). In this study, an *E. coli* type K-12 phosphoglucose isomerase (Δpgi) mutant strain was transformed with a plasmid coding for IFNγ and tested for its expression capabilities, plasmid copy number and mRNA coding for IFNγ number. In addition, a detailed network comprising 100 metabolites and 114 reactions of the central carbon metabolism of this strain was constructed. Then, elementary mode analysis was performed to check flux efficiency from *pgi* mutation and it was predicted that the mutant would have a higher efficiency towards plasmid and protein synthesis. This hypothesis was corroborated experimentally as there was a 3.0-fold increase in IFNγ in the Δpgi mutant.

At the same time, in a work done by Vieira (2015), a generic pipeline for enumeration of minimal cut sets in stoichiometric metabolic models (based on MCSEnumerator algorithm (Kamp & Klamt, 2014)) was implemented and validated. These types of algorithms are relevant in this work as they enable the possibility to enumerate knockouts in a more efficient and simplified manner. Without these methods, enumerating MCSs would be very demanding and would require high computation power even to compute lower sized knockout solutions.

Bearing this in mind and combining these two works together, the main objective of this thesis is to apply minimal cut sets algorithm to find solutions for the optimal and efficient plasmid

and recombinant protein production. In the present work, IFNγ was used as the recombinant protein to apply the MCS algorithm developed by Vieira (2015). IFNγ is a dimerized soluble cytokine that plays a critical role in innate and adaptive immunity against mainly viral infections. Besides its ability to inhibit viral replication, IFNγ plays a big role in the immune system with its immunostimulatory and immunomodulatory effects. As a therapeutic agent, this protein can be used to treat chronic granulomatous disease, that is a condition in which cells of the immunity system have difficulty forming superoxide radical to kill certain pathogens.

Furthermore, the central carbon metabolic network developed by Pandey (2018) and a genome-scale *E. coli* K-12 metabolic network were used. To these models, plasmid, recombinant protein and resistance marker production reactions were formulated in four different ways and added. Then, different MCS enumeration problem formulations were constructed and applied to these models.

Lastly, all the results from both models were analysed with the objectives of: **(1)** corroborating the findings from Pandey *et al.* (2018) that *E. coli pgi* mutant increases plasmid and recombinant protein production flux efficiency; and **(2)** identifying a possible new knockout or set of new knockouts strategies that would lead to a more optimal and efficient plasmid and/or recombinant protein production and that are biologically relevant and feasible.

Chapter 2

Materials and Methods

In this Chapter, the methodology used in the practical part of this work is described. Emphasis on the framework and mathematical algorithms is added.

2.1 Metabolic Models

All metabolic models that were used to perform simulations and their characteristics are detailed in the following sections.

2.1.1 Central Metabolism Model

The Central Metabolism Model (CMM) used throughout this work has its foundation in a model constructed by Pandey *et al.*, 2018. It is a small detailed network of the *E.coli* central carbon metabolic pathway. This network comprises 100 metabolites and 114 reactions (Appendix A), where 9 are exchange and 17 are reversible (the remainder are internal and irreversible reactions). From the list of metabolites, only seven are considered external, those being glucose, ammonium, phosphate, oxygen, carbon dioxide, ethanol and acetate. Glucose is considered the sole carbon source whose cell uptake is done via the phosphotransferase system (PTS). Simultaneously, ethanol and acetate are the overflow metabolites secreted that are produced by the cell to balance the NADH/NAD^+ pool and obtain extra ATP under high carbon source uptake rate or low oxygen availability, which is normal during batch growth. Once secreted, these metabolites can be consumed back, thus the glyoxylate cycle, gluconeogenesis and Entner-Doudoroff (ED) pathways were added. The synthesis of nucleotides and amino acids, essential to biomass production, were also included separately. As for transhydrogenase activity, *E.coli* is known for having two transhydrogenases, that are represented in this network by two reactions with a cost of 0.25 mole ATP per mole of produced NADH. Regarding energy balance, on the one hand, maintenance energy requirements were addressed by including an ATP hydrolysis reaction. On the other hand, for ATP regeneration via oxidative phosphorylation, both NADH and FADH were considered separately with a yield of 2 and 1 mole of ATP on one mole of NADH and FADH, respectively. Furthermore, biomass pseudo reaction was constructed with amino acids, nucleotides, lipids and other requirements. Recombinant proteins and plasmids

were synthesized using amino acids and nucleotides, respectively, accounting energy expenditures (further details in Section 2.1.3 - Model Formulations).

2.1.2 Genome-scale Model

The Genome-scale model (GSM) used throughout this work was iJO1366, whose reconstruction was done by Orth *et al.*, 2011. It is an extremely detailed network that is representative of the *E. coli* K-12 MG1655 metabolism and that was expanded from a previous model, the *i*AF1260. The updated version of this network was obtained from *BiGG Models* database and presently accounts for 1367 associated genes, 2585 metabolic reactions and 1805 metabolites. Unlike CMM, these metabolites and reactions can be compartmentalized in cytoplasmic, periplasmic or extracellular, adding another level of complexity. Additionally, this model has 39 subsystems, from which alanine and membrane lipid metabolism to glycolysis/gluconeogenesis and tricarboxylic acid (TCA) cycle, are just some examples. Recombinant proteins and plasmids synthesis were added to the model using amino acids and nucleotides, respectively, and accounting energy expenditures that are further detailed in the following Section 2.1.3.

2.1.3 Model Formulations

In model formulations, the objective was to construct stoichiometric reactions for the synthesis of a plasmid, its resistance marker and a recombinant protein. Additionally, different protein producing metabolic networks and ways to formulate the enumeration problems were developed.

Recombinant Protein

Regarding the recombinant protein synthesis, the selected model protein for this work was the human interferon gamma (IFNγ) as studied by Pandey *et al.*, 2018. This synthesis reaction was included by quantifying the per mole amino acid requirement for the His-tagged IFNγ (Appendix B) and assuming 4.3 ATPs per peptide bond as it is, approximately, the necessary energy to condensate two amino acids. Protein primary sequence and composition is available at *NCBI* database reference sequence number NP_000610.2 (Interferon gamma precursor [homo sapiens]) and to this sequence, a 6 histidines His-tag was added to perform stoichiometric computations, consistent with the protein produced experimentally by Pandey *et al.*(2018).

Plasmid

For plasmid synthesis, the selected model plasmid for this work was the pET28a vector system from *Novagen* as used by Pandey *et al.*, 2018. This synthesis reaction was included by quantifying the per mole deoxyribonucleotide triphosphate (dNTP) requirement for the pET28a-IFNγ system (considering the His-tag) (Appendix B). The necessary energy to condensate two dNTPs was assumed to be approximately 1.36 ATPs per nucleotide bond. Plasmid primary sequence and composition is available at *Addgene* database and to this sequence, a nucleotidic

IFNγ sequence that is available at *NCBI* database accession reference AB451324.1 was added to perform stoichiometric computations.

Resistance Marker

A resistance marker synthesis reaction was added based on the plasmid antibiotic resistance. The pET28-a vector system presents a kanamycin resistance marker and thus a reaction was included quantifying the per mole amino acid requirement for the production of the enzyme that confers resistance to kanamycin (aminoglycoside O-phosphotransferase APH(3')-Ia). The energy expenditures were assumed to be 4.3 ATPs per peptide bond and the primary sequence and composition of this phosphotransferase was obtained from *NCBI* database reference sequence number WP_000018329 (aminoglycoside O-phosphotransferase APH(3')-Ia [Bacteria] (kanR)).

Model Configurations

In addition to the metabolic reactions present in the models, different ways to balance the equations of plasmid and/or IFNγ synthesis were considered, giving rise to different ways to represent the *E. coli* K12 system. In total 4 different balance equation formulations were created and all the changes to both models (CMM and GSM) were done in MATLAB using COBRA Toolbox.

The base model is the simplest and comprises only a reaction to account for plasmid synthesis. It does not contain in its stoichiometric matrix any information regarding IFNγ and phosphotransferase. Thus, this model is built on an assumption that plasmid and recombinant protein production are directly proportional, meaning that the more plasmids there are, the more recombinant proteins will be translated from those plasmids at a given time. Equation 2.1 represents, without adequate stoichiometry, the reaction added to this model.

$$\sum \text{pET28a-IFN}\gamma \text{ dNTPs} + \text{ATP} \rightarrow \text{Plasmid} + \text{ADP} + \text{Pi} \tag{2.1}$$

Moreover, another level of detail was added to the previous base model. A IFNγ synthesis reaction was added and is independent from the plasmid reaction. This model treats both plasmid and recombinant protein as uncorrelated entities. From this model, it can be interesting to visualize the flux to one product or another since their monomers' origin is metabolically distinct. The following Equation 2.2 represents the new reaction added.

$$\sum \text{IFN}\gamma \text{ amino acids} + \text{ATP} \rightarrow \text{IFN}\gamma + \text{ADP} + \text{Pi} \tag{2.2}$$

For the third model, a resistance marker synthesis reaction was joined to the previous model. This reaction is independent from the plasmid and IFNγ reaction, only relying on its primary amino acid sequence as precursors. All the entities are uncorrelated and independent from each other. From this model, it can be interesting to investigate how the system behaves and what

options are available when constraints are imposed.

$$\sum \text{Phosphotransferase amino acids} + \text{ATP} \rightarrow \text{Phosphotransferase} + \text{ADP} + \text{Pi} \quad (2.3)$$

Regarding the fourth and last model, a different approach was investigated. In this model, all the reactions are correlated with each other, meaning that IFNγ and phosphotransferase production are directly dependent on plasmid availability. In turn, plasmid availability is dependent on dNTPs as described in Equation 2.1. In addition, IFNγ and phosphotransferase depend on amino acids availability as described in Equations 2.2 and 2.3, respectively. Hence, for this purpose, and since there is not available information on ratios such as recombinant protein formed per plasmid, it was assumed that 1 mole of plasmids would give rise to 1 mole of IFNγ and 1 mole of phosphotransferase, as represented by Equation 2.4.

$$1 \text{ Plasmid pET28a} \rightarrow 1 \text{ IFN}\gamma + 1 \text{ Phosphotransferase} \quad (2.4)$$

From a biological standpoint this is the closest to reality since there is a correlation between products. However, from a computational point of view, it may not work as intended as it is metabolically heavy for the network to mathematically allocate all these fluxes while maintaining biomass growth. Table 2.1 summarizes all the models previously described, as well as a key that will be used throughout this work to simplify the analysis when referring to each model.

Table 2.1: Model configuration key and main aspects summary based on the previously described balance equations.

Model	**Equations**	**Comment**
A	Eq. 2.1	Plasmid production. Base model with simplest configuration.
B	Eq. 2.1 Eq. 2.2	Plasmid and IFNγ production. Independent reactions.
C	Eq. 2.1 Eq. 2.2 Eq. 2.3	Plasmid, IFNγ and phosphotransferase production. Independent reactions.
D	Eq. 2.4	IFNγ and phosphotransferase production dependent on plasmid availability.

Problem Configurations

In addition to the distinct model constructions, different enumeration problem configurations were developed based on yield constraints. In total, four different configurations were implemented (Table 2.2).

The first constraint to be tested was to block solutions where product per biomass yield was below a certain threshold. These products may be the plasmid, IFNγ and phosphotransferase, depending on which model is used. For instance, for model A it is only possible to perform simulations blocking low plasmid per biomass yield. However, model B simulations may have, in addition to plasmid, IFNγ per biomass yield constraints. These constraints are treated and computed individually, hence one simulation per product yield constraint is performed. Similarly, in the second set of constraints, product per biomass yield is considered. However, in this configuration, simulations are run considering all possible constraints at the same time (instead of individually). For instance, in model C, one simulation is run where it will be considered a plasmid, IFNγ and phosphotransferase per biomass yield threshold constraint simultaneously.

Furthermore, the third and fourth constraints are similar to the first and second, respectively. Instead of considering product per biomass, product per plasmid yield thresholds are applied in the enumeration problem. Table 2.2 summarizes all the configurations previously described as well as a key that will be used throughout this work to simplify the analysis when referring to each enumeration problem configuration.

Table 2.2: Problem configuration key and main aspects summary based on the previously described constraints.

Problem	Comment
1	Block low product per biomass yield thresholds individually. Products may be plasmid (P), recombinant protein (R) and resistance marker (M).
2	Block low product per biomass yield thresholds simultaneously.
3	Block low product per plasmid yield thresholds individually. Products may be recombinant protein (R) and resistance marker (M).
4	Block low product per plasmid yield thresholds simultaneously.

The problem and model configuration keys will be used together throughout the rest of this work to simplify the analysis and discussion. For instance, when referring to results of CMM_A1M, one is referring to a simulation performed on the a CMM model that only has a plasmid production reaction (model A) and whose enumeration problem was constrained to block low phosphotransferase per biomass yield (1 means product per biomass yield and M refers to the product, in this case the resistance marker). Another example, GSM_C4 is referred to the GSM model that has the 3 individual reactions (model C) and whose enumeration problem was constrained to block low product per plasmid yield simultaneously (in this case, IFNγ and phosphotransferase per plasmid yield, at the same time).

2.2 Cellular Constraints

To solve MCS and FBA problems, biological or physiochemical cellular constraints need to be added to limit the solution space to achieve desirable phenotypes. As the main objective was to evaluate the system behaviour, most cellular constraints are not extremely strict. Glucose maximum uptake rate was set to 1000 $mmol/g \cdot h$ as well as the maximum oxygen consumption rate. These bounds do not have any physiological and biological meaning. However, this way, the model has more freedom to use its main substrate sources and it is possible to evaluate whether producing a by-product (recombinant protein, for instance) is viable with cell growth. Moreover, the upper and lower bounds on cellular maintenance energy (ATPM reaction) were left at the empirical default of 8.39 $mmol/g \cdot h$ (Orth *et al.*, 2010). In addition to the previous constraints, a minimum biomass and product per substrate yield threshold were added. Not desiring to constraint too much the problem formulation, these values were both set to 0.0001. These cellular constraints were maintained in all simulations in this work and are presented in summary in Table 2.3.

Table 2.3: Cellular constraints applied to all the simulations and models used throughout this work.

Constraint	**Value** ($mmol/g \cdot h$)	**Comment**
Glucose	1000	Glucose maximum uptake rate set to not constraint too much the system
Oxygen	1000	Oxygen maximum uptake rate set to not constraint too much the system
Maintenance	8.39	ATPM reaction set to empirical value as requirement cell maintenance.
Biomass	0.0001	Biomass reaction minimum threshold
Product/Glucose	0.0001	Product per substrate minimum yield threshold

To perform FBA and pFBA simulations, the maximization of biomass growth was the elected objective function as it is the most commonly used biological optimization goal.

2.3 Enumeration Algorithm

To compute the MCS/cMCS enumeration problems, a method developed by Vieira (2015) was provided. In this work, Vieira implemented in *Java* programming language a library containing routines for MCS enumeration that can be used from small networks to genome-scale metabolic models. In this context, the pipeline constructed by Vieira and incorporating four main steps is depicted in Figure 2.1.

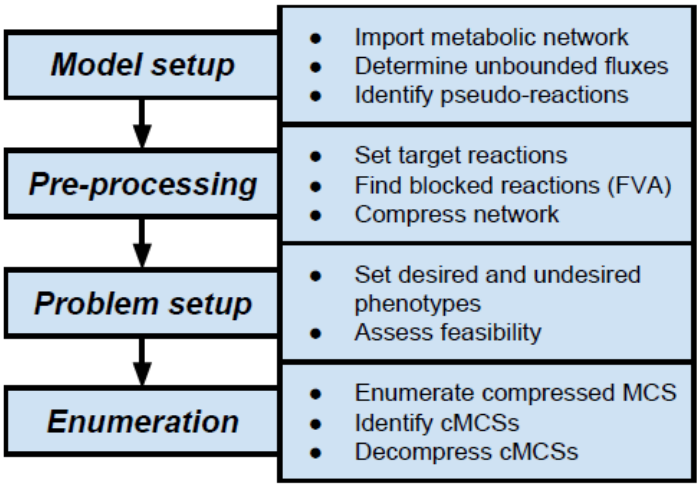

Figure 2.1: Generic pipeline for enumerating MCSs featured in Vieira (2015) work (From Vieira, 2015).

- **Model setup:** In this step, the model is imported and unbounded fluxes are removed to improve numerical stability. In addition, pseudo-reactions are identified as they will not be part of the solutions.

- **Pre-processing:** This steps aims to improve computational speed by reducing the network complexity. It is accomplished by removing blocked reactions that are found through flux variability analysis (FVA) and network compression by lumping correlated reactions. This compression is based on the enzyme subset concept from nullspace analysis, where each enzyme subset is considered a single reaction.

- **Problem setup:** In this step, the enumeration problem is assembled and validated before continuing. A group of desired and undesired phenotypes is constructed based on flux bounds (acting as capacity constraints) and yield constraints (that forces the ratio between two fluxes to a threshold). After defining the phenotypic space, problem feasibility is assessed.

- **Enumeration:** In this last step, the proper formulation is built in the pre-processed model and solved. The K-shortest algorithm is used to compute the EFMs (Figueiredo *et al.*, 2009) and the MCSEnumerator algorithm to enumerate the MCSs (Kamp & Klamt, 2014). Then, the MCSs are checked for feasibility in the desired conditions. Finally, solutions are decompressed with simple combinatorics and MCSs that do not belong to the desired phenotypic space are discarded, leaving only cMCSs.

The provided problem formulation script was modified, using *Eclipse* software, to accommodate the desirable phenotypes for this work described in the previous sections.

2.4 Statistical Methods

2.4.1 Principal Component Analysis

Principal component analysis (PCA) is an unsupervised learning method that aims to reduce the high dimensionality of a dataset while retaining its variation, patterns and trends. This dimensionality reduction is achieved by defining new variables that are a linear combination of the original ones and, geometrically, are the orthogonal projection - the principal components (PC). It is done as such that the first PC has the largest possible variance (accounting for as much of the variability in the data) and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components (Ringnér, 2008).

Firstly, the data may require some processing, such as its centralization and standardization, taking into consideration that PCA is sensitive to the relative scaling of the original variables. Then, a correlation matrix is computed, containing the correlation values between all pairs of variables. From this matrix, the eigenvectors and eigenvalues can be extracted, describing the directions of patterns in data and the variance explained by these directions, respectively. Eigenvectors correspond then to the principal components and the eigenvalues to the variance each component explains (Smith, 2002; Ringnér, 2008).

PCA is useful and very common in biology as it helps reduce the high dimensionality in, for instance, NGS data where the number of samples is significantly lower than the number of features (genes or transcripts). The samples can then be plotted according to their projection onto each of the components, allowing the visualization of possible patterns and groups contained in it.

In this work, PCA was used with the purpose of searching for patterns in reaction fluxes and to group similar solutions, ultimately to reduce the solution pool size. Principal component analysis was implemented in *R* using the function *PCA* from the *FactoMineR* package (Husson *et al.*, 2017).

2.4.2 Cluster Analysis

Cluster analysis (or clustering) is an unsupervised learning methodology which means there are no predefined data labels or classes. The main goal of these methods is to group a set of objects in such way that objects in the same group (called clusters) are more similar to each other than to those in other clusters. The similarity/dissimilarity is a key component in clustering as it is the main controlling factor when grouping data and it is typically expressed in terms of distance. For such calculations, a distance metric is required and it is chosen according to the features and type of data available. The most popular metrics are the Manhattan and Euclidean distances that calculate the distance between data points as given by Equations 2.5 and 2.6 , respectively (Rokach & Maimon, 2005).

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + ... + |x_{ip} - x_{jp}| \tag{2.5}$$

$$d(i,j) = \sqrt{\left|x_{i1} - x_{j1}\right|^2 + \left|x_{i2} - x_{j2}\right|^2 + ... + \left|x_{ip} - x_{jp}\right|^2} \tag{2.6}$$

The appropriate clustering algorithm and parameter settings (including distance function, a density threshold or the number of expected clusters) depend on the data set and intended use. As such, cluster analysis is rather an iterative process of knowledge discovery and it is often necessary to modify data preprocessing and model parameters until the results achieve the desired properties.

Hierarchical Cluster Analysis

Hierarchical cluster analysis (HCA) is a method which seeks to build a hierarchy where clusters have subclusters that are organized in a tree, and each node (clusters) is the union of its children (subclusters). It can fall into two categories: the agglomerative and divisive approach.

The agglomerative clustering methods are the most used and work in a "bottom-up" manner. That is, each object is initially considered a single-element cluster (leafs) and at each step of the algorithm the two most similar clusters are merged into one single big cluster (nodes). This procedure is iterated until a stopping criteria is met or all elements are joined in one single cluster (root). This allows the construction of a nested grouping of patterns, usually represented in a dendrogram (Gan *et al.*, 2007).

On the other hand, divisive clustering methods are essentially the inverse process of an agglomerative technique and work in a "top-down" manner. In this case, at the start, all objects are included in one single cluster (root). At each step of iteration, the most heterogeneous cluster is divided into two (nodes). This process is iterated until each object has its own single-element cluster (leafs) (Gan *et al.*, 2007).

In this type of clustering the clusters of a higher hierarchy level encompass all the objects that belong to the merged clusters from the lower level, which means that when an object is assigned to a certain cluster it is not possible to reassigned to another cluster. In Figure 2.2 the two HCA approaches described above are represented (Rokach & Maimon, 2005).

To compute the dissimilarity between two clusters of observations there are three popular methods:

- **Single linkage:** It computes all pairwise dissimilarities between elements in cluster A and elements in cluster B, and considers the smallest of these dissimilarities as a linkage criteria.

$$min\{d(x,y) : x \in A,\ y \in B\} \tag{2.7}$$

- **Complete linkage:** It computes all pairwise dissimilarities between elements in cluster A and elements in cluster B, and considers the maximum value of these dissimilarities as the distance between the two clusters.

$$max\{d(x,y) : x \in A,\ y \in B\} \tag{2.8}$$

- **Average linkage:** It computes all pairwise dissimilarities between elements in cluster A and elements in cluster B, and considers the average of these dissimilarities as the distance between the two clusters. It is an in between to the last two methods.

$$\frac{1}{|A| \cdot |B|} \sum_{x \in A} \sum_{y \in B} d(x, y) \tag{2.9}$$

Hierarchical algorithms are suitable for dataset with arbitrary shape and is advantageous as it outputs a hierarchy, that is a structure that can be more informative cluster-wise when comparison to unstructured sets of flat clusters returned in other algorithms. However, it presents high inability to adjust decisions, given the impossibility of reassigning the objects to different clusters after their assignment. It is also very sensitive to outliers and is not suitable for very large datasets (Santini, 2016).

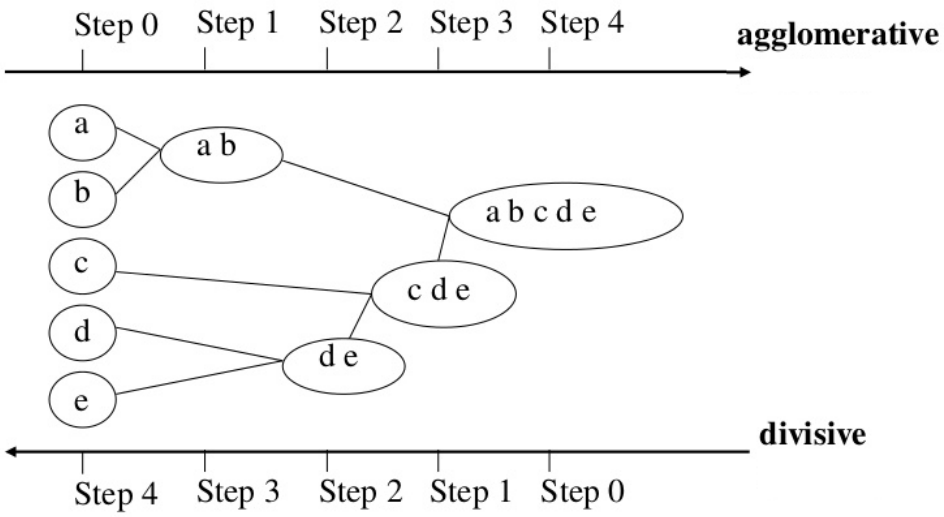

Figure 2.2: Representation of the agglomerative and the divisive HCA approach.

2.5 Tools and Software

2.5.1 The R Programming Language

All the analytical work described throughout this thesis was implemented in *R* programming language, a free software environment for statistical computing and graphics. *R* provides a grand variety of not only statistical methods but also good graphical visualization and can be easily extended using a wide range of available packages, thereby facilitating data analysis and visualization (R Development Core Team, 2011). In order to run *R*, a user-friendly environment, *RStudio*, that is an integrated development environment (IDE) was used.

The main majority of graphics and data visualization displayed throughout this work were generated through the *ggplot* function from the *ggplot2* package available for *R* (Wickham, 2009).

2.5.2 MATLAB

MATrix LABoratory (MATLAB) is a computing environment and programming language for algorithm development, data analysis, visualization and numerical computation. It allows matrix calculations, plotting of functions and data, creation of graphical user interfaces (GUI) and algorithms that can be implemented or used from a vast library of pre-built toolboxes. In addition, it is possible to interface with programs developed in different languages (C/C++, Java®, .NET, Python, SQL, Hadoop and Microsoft® Excel®) which makes it possible to harness the unique strengths of each language for various purposes.

Throughout this work, MATLAB was used to manipulate and construct all the metabolic models that were simulated. All these changes were achieved using the COBRA Toolbox (Heirendt *et al.*, 2018).

2.5.2.1 The COBRA Toolbox

The COnstraint-Based Reconstruction and Analysis (COBRA) Toolbox is a set of methods and utilities integrated as a form of software package to provide easy access to core COBRA methodologies. It provides methods for quantitative prediction of cellular phenotype and multi-cellular biochemical networks with constraint-based modelling. It implements an extended and comprehensive collection of modelling methods, from reconstruction and model generation to prediction and analysis methods. The openCOBRA project has been developing, starting with tools for MATLAB that are currently at their third version (The COBRA Toolboox v3.0) and that have been expanding to Python (COBRApy) and Julia (COBRA.jl) modules (Heirendt *et al.*, 2018).

The COBRA Toolbox for MATLAB was used to deal with the model reconstruction and refinement. This toolbox offers the processing of SBML (Systems Biology Markup Language) files of metabolic networks, as well as the capability of reading and writing. Model manipulation functions such as *addReaction* and *addMetabolite* were commonly used to add information to the network regarding the synthesis of plasmids or recombinant proteins. Some functions were called to export the new models; however, this toolbox was not used to apply any phenotype prediction or analysis algorithm.

2.5.3 The Java Programming Language

In order to run the *Java* scripts to perform flux distributions simulations (pFBA and FVA) and apply the necessary problem formulation modifications, a user-friendly integrated development environment (IDE), *Eclipse*, was used. This software contains a base workspace with an extensive plug-in system for further customization. It is primarily used to develop *Java* applications but may be also used with other programming languages. For this work, *Eclipse Oxygen* version, released on June 2017, was used.

Chapter 3

Results and Discussion

In this Chapter, the main results generated with the different models are presented, analyzed and discussed. Regarding the Central Metabolism Models (CMM), an exploratory data analysis was performed with the aim of finding a small number of solutions for a further detailed network analysis. For the Genome-scale Model (GSM), only an exploratory data analysis was performed in order to evaluate differences comparing to the smaller model.

3.1 Central Metabolism Model

3.1.1 Data Processing

Data were generated for each enumeration problem (combinatorial model and problem configurations) as previously described in Chapter 2. A maximum knockout size of 5 was allowed and all the solutions were stored as sets of strings encoding reactions. For each generated solution, a pFBA flux distribution was computed and stored in a matrix where each row is a solution and each column encodes a reaction. Hence, each matrix entry represents a flux value for a given reaction in a particular set of knockouts (solution).

Before analysing the data, a pre-processing step was performed in order to help reducing data high dimensionality. In this step, some solutions were filtered based on their set of knockouts. On one hand, solutions that were biologically irrelevant were removed. These are solutions that comprise one or more reactions regarding: (1) **production**, such as biomass, plasmid, recombinant protein and resistance marker reactions that are the objective of this work making their removal meaningless; (2) **energy**, such as ATP maintenance and synthesis reactions that are essential to cell survival; and (3) **transport** such as glucose exchange reaction that is assured by the PTS system and are also vital to cells. Table 3.1 summarizes all the reactions that were targets of this filtration step.

On the other hand, solutions that were computationally irrelevant were removed. These were selected based on biomass-product coupled yields (BPCY) and combined reaction flux values, depending on each model and formulation available. For instance, it can be considered that, solutions whose BPCY was above zero or solutions that present a flux different from zero in

plasmid and recombinant protein reactions, at the same time, are the ones to be kept for further analysis.

These filtration steps were applied to each model and formulation. After processing, each formulation data corresponding to a model was concatenated and analyzed simultaneously.

Table 3.1: Biologically relevant reactions that were filtered in the CMM data processing step. Solutions (set of knockouts) that comprised one or more of these reactions were removed from further analysis.

Reaction	Description
R_BiomassProduction	Biomass production reaction
R_PlasmidProduction	Plasmid production reaction
R_RecProduction	Recombinant protein production reaction
R_ResProduction	Resistance marker production reaction
R_ATPM	ATP Maintenance reaction
R_ATPS1	ATP Synthesis reaction (NADH related)
R_ATPS2	ATP Synthesis reaction ($FADH_2$ related)
R_PTS	Glucose transport system

3.1.2 Exploratory Data Analysis

In order to understand the main characteristics of the data, some exploratory data analysis methods were applied. A Principal Component Analysis (PCA) was performed after data filtration and standardization with the objective of evaluating the main source of data variation. In addition, a hierarchical cluster analysis (HCA) was performed with the aim of reducing the solution pool by grouping solutions that present different sets of knockouts reactions but show similar phenotypes. These methods were applied on the pFBA flux distribution data.

In the next Sections, these methods' results will be presented and discussed for all the four models, and each model with all its possible problem configurations.

3.1.2.1 Model A

Model A takes only into consideration the plasmid production reaction. Consequently, there is only one way to compose the enumeration problem, which is by constraining low plasmid production per biomass yields (formulation 1P). From the initial 723 different solutions obtained for this problem, only 8.2% remained for further analysis after the processing step. Table 3.2 summarizes the number of suggested knockouts in a solution (MCS Size) and the amount of solutions that have that size.

Table 3.2: Model CMM_A Summary of the suggested knockout number in a solution (MCS Size) and the amount of solutions that have that size (# MCS), corresponding to each formulation before (pre-) and after (post-) processing steps.

	Formulation	MCS Size	# MCS
Pre-Processing	1P	1	0 (0%)
		2	2 (0.3%)
		3	1 (0.2%)
		4	264 (36.5%)
		5	456 (63.0%)
Total			723
Post-Processing	1P	1	0 (0%)
		2	2 (3.4%)
		3	1 (1.7%)
		4	4 (6.8%)
		5	52 (88.1%)
Total			59

It is noticeable that most of the solutions in the pool suggests a four or five set of knockout reactions. Smaller solutions account for less than 1 % of the pre-processed data and there are no MCSs with only one reaction. After the filtration step, small solutions number remained the same, and solutions with larger knockout pools were heavily reduced (a 98.5% and 88.6% reduction regarding MCSs with a size of 4 and 5 knockouts, respectively).

Moreover, to better visualize and analyze the PCA results, a scree plot was computed showing the variance explained by each principal components until the tenth component. In addition, a correlation circle accounting variables (network reactions) and a graph of individuals (solutions) was computed. The individuals are represented by their projections and the variables are represented by their correlations. Lastly, a HCA was performed to try to cluster solutions. Since the resulting tree is too large, only a specific sub-tree will be shown in the results but the full dendrogram is in Appendix C.

A scree plot is a useful visual tool for determining an appropriate number of principal components that explain the most variability in the data. Figure 3.1 plot shows that five components explain approximately 98.8% variance in these data, *i.e.*, the majority of the data can be reduced to this amount of dimensions without compromising on explained variance and losing important information. There is no defined objective way to decide the number of required components as this depends on each individual dataset and its application. However, in practice, one aims to analyze the first few components in order to find interesting patterns in the data. Regarding model A data, two principal components were chosen to be analyzed as they account for a reasonable fraction of the total variance - around 86.2% cumulative explained variance percentage.

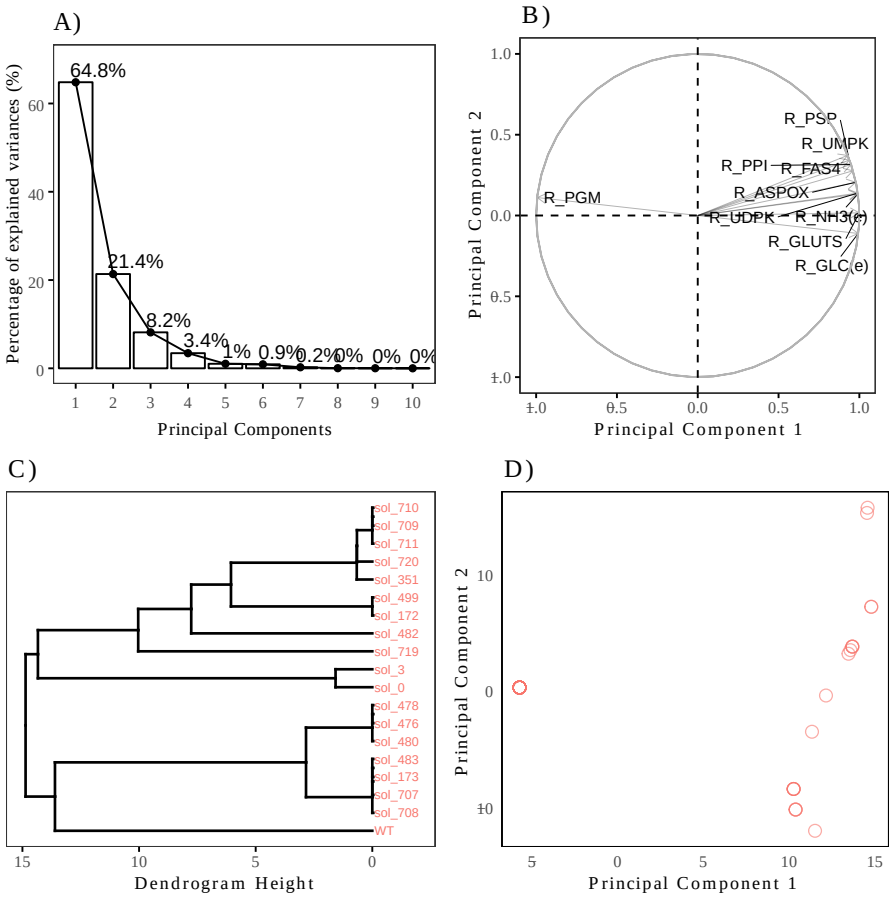

Figure 3.1: Model CMM_A Exploratory data analysis results: **A) Scree plot** percentage of explained variances (%) in each principal component (up to a total of ten PCs); **B) Correlation circle** correlation between the top 10 variables contributing to the PCs and the first and second principal components; **C) Dendrogram** hierarchical cluster analysis performed using single linkage method and euclidean distance metric. The sub-tree was obtained by a cut done at a dendrogram height equal to 15; **D) Individuals graph** data projection coordinates in the first two principal components: ○ CMM_A1P.

A correlation plot gives the variables direction vectors and helps describing the strength of relationship between two variables. A correlation coefficient ranges from -1 to +1, where +1 indicates a perfect positive linear relationship, and -1 a perfect negative linear relationship. Variables which have little contribution to a direction have almost zero weight. Drawing this plot may contribute to understand, in general, which variables have positive and negative impact on the principal components. Strongly related variables, will be positively correlated, and in this plot, will appear near each other. Meanwhile, negatively correlated variables will appear diagonally opposite to each other. If two variables are unrelated, they will appear orthogonal to each other. Additionally, the length of an arrow represents how well it explains the distribution of the data. A variable with a long arrow means it is better represented in the principal components. Figure 3.1 correlation circle shows the top ten contributing variables to the fist and second dimensions. From this plot, all arrows have a similar length, so the parameters contribute equally. Moreover, nine out of ten variables are strongly positively correlated, whereas *R_PGM* is negatively correlated. The latter, corresponds to a reaction in gluconeogenesis where glucose-6-phosphate is transformed back to glucose for energy reservation. In this way, it is valid that this variable is negatively correlated as the remaining reactions are mostly essential to nucleotide

synthesis, where there is a high energy and substrate consumption, in contrast to *R_PGM* whose objective is completely the opposite. Furthermore, some of the positively correlated reactions concern amino acid synthesis such as *R_PSP* (serine synthesis) and *R_ASPOX* (aspartate synthesis). Even though in this model there is not a reaction accounting for recombinant protein production, these amino acids are essential for the pseudo biomass reaction. Additionally, these are also fundamental early precursors in nucleotide synthesis for the plasmid production (serine is involved in MetTHF synthesis and aspartate in PRAIC synthesis). Overall, these top contributing variables show that there is high variation in reactions concerning nucleotide synthesis which is consequently related to plasmid production.

The individuals graph, also known as score plot, is a projection of the data scores into principal components and it is used for finding and interpreting relationships between individuals/observations. It can be used to assess the data structure and detect clusters, outliers, and trends. Groupings of data on the plot may indicate two or more separate distributions in the data. In this model, there is only one possible enumeration problem and, thus PCA is not as helpful in pattern visualization as there are not any separations to be made. Although PCA does not compute which solutions belong together, since to do that clustering methods are needed, it can still give a good visualization of how individual solutions are grouped. For instance, from this score plot it is possible to say that an amount of solutions are very closely grouped (highlighted by the strong pink coloured circle in the left, resulting of solution overlapping) and that contribute exclusively to the first dimension. These solutions may be a possible cluster that could reduce the solution pool as they may represent the same phenotype. In this group, most of the solutions have a MCS length of 5, where 4 suggested knockouts remain the same (*R_TRANSH2*, *R_ACK*, *R_ADH* and *R_SDH*) and the last reaction is different for each solution. These reactions are tightly related to overflow metabolites that are secreted by cells to balance NADH/NAD^+ and obtain ATP (*R_ACK* for acetate and *R_ADH* for ethanol), as well as other cell mechanisms to balance reducing power such as *R_TRANSH2* for NADPH/NADH and *R_SDH* for $FADH_2$. In addition, these solutions have a similar phenotype to a smaller suggested 2 knockout solutions (*R_TRANSH2* and *R_PDH*) and, thus it may be an interesting target for a further detailed analysis to study and explain how a similar phenotype is achieved by deleting 2 reactions instead of 5.

A dendrogram is a tree diagram used to illustrate the arrangement of the clusters produced by hierarchical clustering, which is useful to find correlated groups. The horizontal axis of a dendrogram represents the distance or dissimilarity between clusters, whereas the vertical axis represents the objects and clusters. Each joining (fusion) of two clusters is represented on the graph by the splitting of a horizontal line into two horizontal lines. The horizontal position of the split, shown by a short vertical bar, gives the distance (dissimilarity) between the two clusters. Furthermore, cutting a dendrogram at a certain level/height gives a set of clusters. Thus, depending at which height the cut is done, one can have variable cluster numbers. There is no definitive height at which a dendrogram should be cut as the resulting hierarchical structure is context-dependent. Figure 3.1 shows a sub-tree that was obtained by cutting the full dendrogram at a height equal to 15. This tree was computed using single linkage method and euclidean distance as the metric distance. Since the main objective behind our HCA is to find solutions

that are highly similar between each other, these methods seem to be the most appropriate as individuals are grouped by how close their squared distances are. Looking at the full dendrogram (in Appendix C), there are two very distinct groups separated by a high dissimilarity. The top group (represented in Figure 3.1 sub-tree) seems to consist of more distinct clusters, while most of the individuals in the bottom group are all clustered together at the same height. Comparing the PCA with the HCA results, it is possible to corroborate that the group in PCA corresponds to an actual cluster in HCA (bottom group) and, thus the phenotypes are equal in all those solutions. These are also the solutions that are less related to the wild-type (WT) which can be an indicator in a sense that, being the primary focus to search plasmid producing phenotypes, these are the complete opposite of the WT. Overall, it is possible to see patterns of clusters that are based on solutions that are closely related as they share 3 or 4 suggested knockouts in common, only differing in 1 or 2 reactions.

In addition to the previous results, the number of times each reaction appeared in all overall solutions (# KO) was computed. This was done to provide a general idea which reactions are deleted repeatedly, as well as to guide the decision on which solutions to further analyze. The top 10 most targeted reactions for knockouts in solutions concerning model A data are summarized below in Table 3.3.

Table 3.3: Model CMM_A top ten most targeted reactions for knockouts in the overall solutions set (#KO).

Reaction	**# KO**
R_TRANSH2	42
R_SDH	42
R_ACK	40
R_ADH	40
R_PGM	17
R_PYK	16
R_PGDH	12
R_PYC	9
R_G6P1D	5
R_ICL	3

The most targeted reactions, which were suggested for knockout 40 and 42 times, had been previously mentioned in the PCA analysis and are tightly related to reducing power balance and ATP production.

3.1.2.2 Model B

Model B considers the individual plasmid and recombinant protein production. Consequently, there are multiple ways to formulate the enumeration problem - formulations 1P, 1R,

2, 3R and 4. On average, for each formulation, from the initial number of different solutions obtained, only 3.2 % remained for further analysis after processing. As a whole, from the 3649 total solutions, only 115 were left for further analysis, which corresponds to a 96.8 % decrease in total solutions. Table 3.4 summarizes the MCS size and the amount of solutions to that corresponding size (# MCS) for enumeration problems performed on this model.

As seen previously, most solutions in the pool comprise sets of four or five knockout reactions, whereas smaller solutions account for less than 1% of the pre-processed data. No MCSs were found with only one reaction. The pattern repeats itself, as most of the smaller solutions remain after processing, with larger knockout sets being heavily reduced (an average processing reduction of 98.9% and 97.6% for MCSs with sizes of 4 and 5 knockouts, respectively). In addition, there are formulations that do not present any MCSs with 3 or less knockouts - formulations 1R, 3R and 4.

To better understand and visualize these differences and results, the PCAs and HCAs performed are shown below in Figure 3.2.

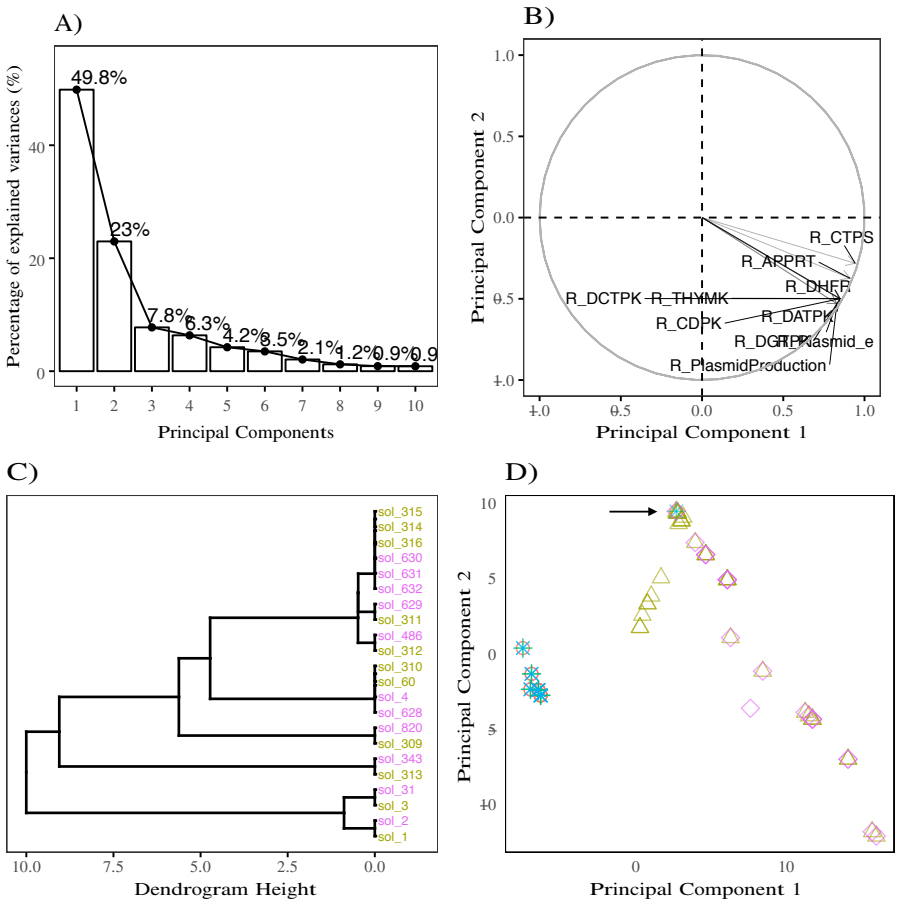

Figure 3.2: Model CMM_B Exploratory data analysis results: **A) Scree plot** percentage of explained variances (%) in each principal component (up to a total of ten PCs); **B) Correlation circle** correlation between the top 10 variables contributing to the PCs and the first and second principal components; **C) Dendrogram** hierarchical cluster analysis performed using single linkage method and euclidean distance metric. The sub-tree was obtained by a cut done at a dendrogram height equal to 10.2; **D) Individuals graph** data projection coordinates in the first two principal components: ◇ CMM_B1P × CMM_B1R △ CMM_B2 ○ CMM_B3R + CMM_B4 .

Table 3.4: Model CMM_B Summary of the suggested knockout number in a solution (MCS Size) and the amount of solutions that have that size (# MCS), corresponding to each formulation before (pre-) and after (post-) processing steps.

	Formulation	MCS Size	# MCS		Formulation	MCS Size	# MCS
Pre-Processing	1P	1	0 (0%)	**Post-Processing**	1P	1	0 (0%)
		2	3 (0.3%)			2	1 (5%)
		3	1 (0.1%)			3	1 (5%)
		4	186 (21.7%)			4	2 (10%)
		5	669 (77.9%)			5	16 (80%)
Total			859	**Total**			20
Pre-Processing	1R	1	0 (0%)	**Post-Processing**	1R	1	0 (0%)
		2	0 (0%)			2	0 (0%)
		3	0 (0%)			3	0 (0%)
		4	128 (20%)			4	2 (10%)
		5	501 (80%)			5	18 (90%)
Total			629	**Total**			20
Pre-Processing	2	1	0 (0%)	**Post-Processing**	2	1	0 (0%)
		2	3 (0.3%)			2	1 (2.8%)
		3	1 (0.1%)			3	1 (2.8%)
		4	196 (21.7%)			4	9 (25.7%)
		5	703 (77.9%)			5	24 (68.7%)
Total			903	**Total**			35
Pre-Processing	3R	1	0 (0%)	**Post-Processing**	3R	1	0 (0%)
		2	0 (0%)			2	0 (0%)
		3	0 (0%)			3	0 (0%)
		4	128 (20%)			4	2 (10%)
		5	501 (80%)			5	18 (90%)
Total			629	**Total**			20
Pre-Processing	4	1	0 (0%)	**Post-Processing**	4	1	0 (0%)
		2	0 (0%)			2	0 (0%)
		3	0 (0%)			3	0 (0%)
		4	128 (20%)			4	2 (10%)
		5	128 (20%)			5	18 (90%)
Total			629	**Total**			20

From the scree plot it is possible to compute that at least seven principal components are necessary to explain approximately 97.0 % variance in these data. In comparison to the previous model, at least two more dimensions are required to achieve almost the same variance percentage, meaning that, by introducing the recombinant protein reaction in the model, more contrast and divergence was included. Concerning model B data, two principal components were chosen to be analyzed as they account for 72.8 % of cumulative explained variance percentage. Although this value is 13.4 % lower than the previous model, it still considers a reasonable amount of explained variance in just two dimensions. This also corroborates that the IFNγ production

reaction introduced more variation in the system.

From the correlation circle, it is possible to visualize that all top ten contributing variables share the same amount of contribution to the components as their arrows present the same length (equal to the correlation circle radius, which is equal to one). Furthermore, all ten variables are in the negative side of component 2 and positive side of component 1 but are strongly positively correlated with each other. Two interesting reactions that contribute to this variance are the ones related to plasmid production (*R_PlasmidProduction and R_Plasmid_e*). By adding the recombinant protein production reaction, it seems that producing a plasmid became extremely variable and perhaps dependant on precursors availability, now that the cell may require amino acids for IFNγ production. The remaining reactions are mostly related to nucleotide synthesis. Four of these account for deoxyribonucleotide triphosphate (dNTPs) synthesis which are the precursors for plasmid production *(R_DCTPK*, *R_THYMK*, *R_DGTPK* and *R_DATPK* that correspond to the dCTP, dTTP, dGTP and dATP synthesis, respectively). The remaining variables concern other precursors necessary for dNTP synthesis. The only outlier is *R_DHFR* that belongs to the one carbon units family but, nevertheless, produces an important compound for nucleotide synthesis reactions (THF). Overall, by introducing the recombinant protein production, all top contributing variables are related with plasmid production and nucleotide synthesis and, thus it is expected that these reactions present a strong positive correlation.

As far as the individuals graph is concerned, this analysis shows that there is a clear separation between most solutions from formulations 1P and 2, in contrast to formulations 1R, 3R and 4. In addition, a point in space is clearly seen that has all possible formulations overlapped (highlighted by the arrow in Figure 3.2). This point naturally corresponds to the WTs for each formulation as it presents always the same phenotype. The data points that are completely on top of each other suggest a very strong grouping of equal phenotypes. In fact, all the solutions for these three enumerations are exactly the same, meaning that they can be treated as one, having a total of 60 different solutions that can be reduced to 20 solutions that explain the exact same phenotype. Most of these solutions identify a reaction that concerns to reducing power (*R_TRANSH2*) in addition to combinations of reactions from the pentose phosphate pathway (PPP) that are knocked out at different stages (*R_6PGDH*, *R_TALA1*, *R_R5P1*, *R_TKT1*, *R_G1D* and *R_GLUCK*). Moreover, regarding the other two formulations, some solutions may be grouped but there is more variety in these formulations. In addition, a few of these solutions are closely related to the WTs phenotype.

From the full HCA dendrogram (in Appendix C) it is possible to corroborate that there is a complete separation based on dissimilarity for the previously mentioned PCA groups. In this case it is harder to find an evident cut-off height that can be helpful to separate different clusters as there are plenty of options at many heights. Nevertheless, it is possible to at least isolate a group as shown in Figure 3.2 sub-tree where a cut-off of 10.2 was applied. It is also available to see which solutions are closer to the WT phenotype and which ones are not. Overall, this HCA is helpful to visualize the solutions group separation as well as understand that the inclusion of the recombinant protein added a level of variation in the system that is shown by the new multiple ways to cluster all the solutions.

Moreover, the top ten targeted reactions for suggested knockouts in model B data solution set are given below by Table 3.5.

Table 3.5: Model CMM_B top ten most targeted reactions for knockouts in the overall solutions set (#KO).

Reaction	**# KO**
R_TRANSH2	77
R_PGM	60
R_6PGDH	58
R_PYK	38
R_PGI	32
R_PGDH	27
R_G1D // R_GLUCK	23
R_R5PI // R_TALA2 // R_TKT1	22
R_PFK	20
R_PYC	19

Once more *R_TRANSH2* is the most suggested knockout in the solution set. Top KOs are reactions concerning glycolysis and gluconeogenesis. Furthermore, some also concern the PPP and the sole reaction representing the Entner Doudoroff (ED) pathway (*R_PGDH*) is present. In general, all these reactions belong to bacterial primary metabolism and to pathways whose goal is to catabolize glucose to pyruvic acid.

3.1.2.3 Model C

Model C considers the individual plasmid, recombinant protein and resistance marker production, and thus there are multiple ways to formulate the enumeration problem- formulations 1P, 1 R, 1M, 2, 3R, 3M and 4. On average, for each formulation, from the initial number of different solutions, only 2.5 % prevailed for further analysis in the post-processing steps. As a whole, from a total of 2770 solutions, only 76 remained for further analysis, which corresponds approximately to a 97.3 % total solutions decrease. Table 3.6 summarizes the MCS size and the amount of solutions correspondent to that size for all enumeration problems performed on model C.

Table 3.6: Model CMM_C Summary of the suggested knockout number in a solution (MCS Size) and the amount of solutions that have that size (# MCS), corresponding to each formulation before (pre-) and after (post-) processing steps.

	Formulation	MCS Size	# MCS
Pre-Processing	1P	1	0 (0%)
		2	3 (0.4%)
		3	1 (0.1%)
		4	186 (23.5%)
		5	601(76%)
Total			791
Pre-Processing	1R	1	0 (0%)
		2	0 (0%)
		3	0 (0%)
		4	2 (0.7%)
		5	268 (99.3%)
Total			270
Pre-Processing	1M	1	0 (0%)
		2	0 (0%)
		3	0 (0%)
		4	0 (0%)
		5	319 (100%)
Total			319
Pre-Processing	2	1	0 (0%)
		2	3 (0.5%)
		3	1 (0.2%)
		4	196 (34.3%)
		5	371 (65%)
Total			571
Pre-Processing	3R	1	0 (0%)
		2	0 (0%)
		3	0 (0%)
		4	2 (0,7%)
		5	268 (99.3%)
Total			270
Pre-Processing	3M	1	0 (0%)
		2	0 (0%)
		3	0 (0%)
		4	0 (0%)
		5	319 (100%)
Total			319
Pre-Processing	4	1	0 (0%)
		2	0 (0%)
		3	0 (0%)
		4	2 (0.9%)
		5	228 (99.1%)
Total			230

	Formulation	MCS Size	# MCS
Post-Processing	1P	1	0 (0%)
		2	1 (5%)
		3	1 (5%)
		4	2 (10%)
		5	15(80%)
Total			19
Post-Processing	1R	1	0 (0%)
		2	0 (0%)
		3	0 (0%)
		4	2 (100%)
		5	0 (0%)
Total			2
Post-Processing	1M	1	0 (0%)
		2	0 (0%)
		3	0 (0%)
		4	0 (0%)
		5	0 (0%)
Total			0
Post-Processing	2	1	0 (0%)
		2	1 (3%)
		3	1 (3%)
		4	9 (25.7%)
		5	24 (68.3%)
Total			35
Post-Processing	3R	1	0 (0%)
		2	0 (0%)
		3	0 (0%)
		4	2 (100%)
		5	0 (0%)
Total			2
Post-Processing	3M	1	0 (0%)
		2	0 (0%)
		3	0 (0%)
		4	0 (0%)
		5	0(0%)
Total			0
Post-Processing	4	1	0 (0%)
		2	0 (0%)
		3	0 (0%)
		4	0 (0%)
		5	18 (100%)
Total			18

Again, large solutions prevailed over smaller solutions (1, 2 and 3 knockouts), which account for about 1% or less of the pre-processed data. No MCSs with one reaction were identified. Larger knockout solutions were reduced with an average processing reduction of 99.0 % and 97.5 % for MCSs with size of 4 and 5 deletions, respectively. Some formulations do not present solutions with 3 and 4 knockouts or below - formulations 1R, 1M, 3R,3M, 4. Interestingly, some formulations regarding resistance marker production did not survive the processing steps (formulations 1M and 3M). In addition, formulations for the recombinant protein lost 99.3 % of their solutions, where only 2 remained. Overall, with the resistance marker introduction fewer solutions were computed and got through the processing steps.

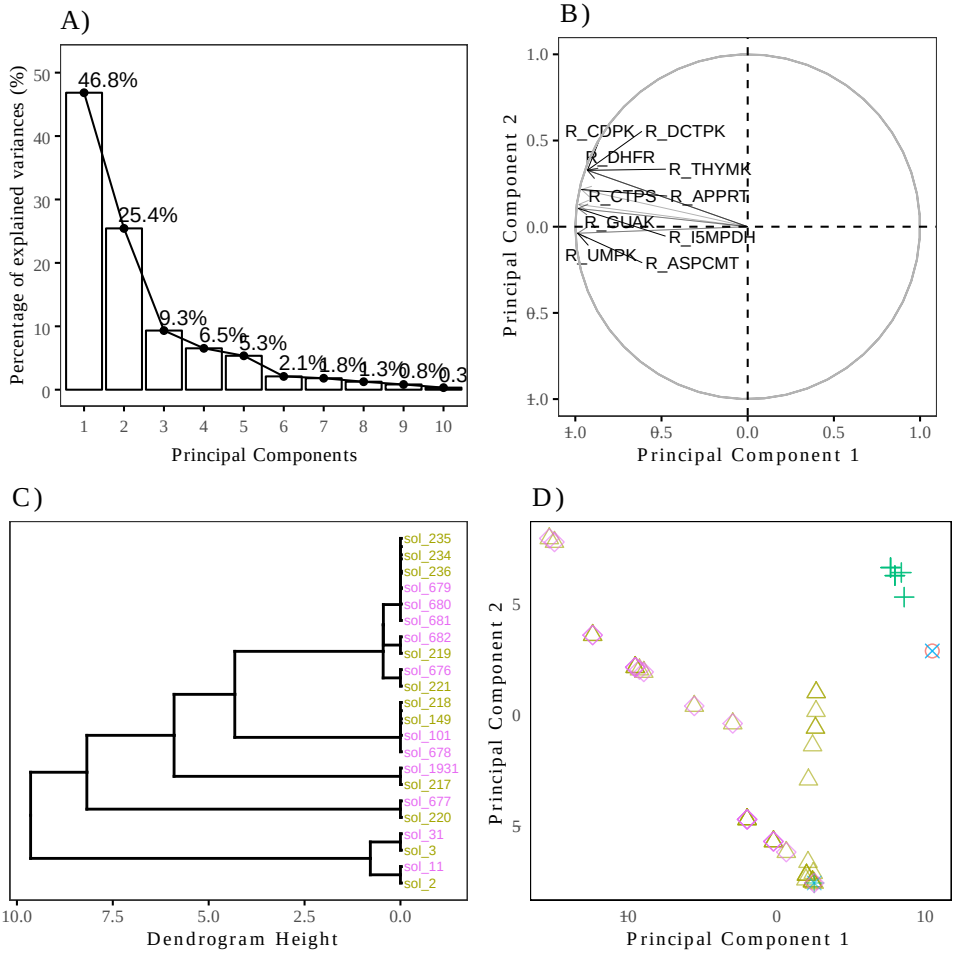

Figure 3.3: Model CMM_C Exploratory data analysis results: **A) Scree plot** percentage of explained variances (%) in each principal component (up to a total of ten PCs); **B) Correlation circle** correlation between the top 10 variables contributing to the PCs and the first and second principal components; **C) Dendrogram** hierarchical cluster analysis performed using single linkage method and euclidean distance metric. The sub-tree was obtained by a cut done at a dendrogram height equal to 10; **D) Individuals graph** data projection coordinates in the first two principal components: ◇ CMM_C1P ✕ CMM_C1R △ CMM_C2 ○ CMM_C3R + CMM_C4 .

The scree plot shows that a minimum of seven principal components are required to explain approximately 97.2 % variance in this data, which is nearly equal to the previous model scree plot. Regarding this model data, two principal components were once more chosen to be analyzed and account for 72.2 % of cumulative explained variance, which is a reasonable amount of

explained variation in a two dimensional space. This value is similar to the previous one, which may be indicative that, by adding the resistance marker production reaction, there was not a major shift and introduction of divergence. This may happen as the resistance marker is essentially another protein to be produced and, thus, the amino acids required for the IFNγ production are the same needed for the resistance marker production, but in different quantities. Comparing to model A, models B and C have a less gap difference as their core dissimilarity relies on one protein production reaction (and not plasmid, where nucleotides are involved instead of amino acids).

Analyzing the correlation circle, it is possible to state that, again, all top ten variables equally contribute to the components as they show equivalent arrow length. All these reactions are in the positive side of the second principal component and the negative side of the first component, and demonstrate a strong positive correlation among each other. In comparison to the previous model, it is interesting to note that, with the addition of the resistance marker production, the reactions regarding plasmid production are no longer on the top contributing variables. Nevertheless, nine out of ten variables belong to the nucleotide synthesis family. Two of these are related with plasmid production precursors (*R_DCTPK* and *R_THYMK* that correspond to dCTP and dTTP synthesis, respectively) and the remaining are related to other precursors necessary for dNTP synthesis. The former being reactions with respect to nucleoside monophosphate (*R_ASPCMT* for UMP synthesis), nucleoside diphosphate (*R_GUAK* and *R_UMPK* for GDP and UDP synthesis, respectively) and nucleoside triphosphate (*R_CDPK* and *R_CTPS* for CTP synthesis). Again, the only outlier corresponds to *R_DHFR* that belongs to the one carbon units family but, nevertheless is important in nucleotide synthesis. Overall, the main differences between this model and the previous rely on the plasmid production reactions. Regardless, on both models, nucleotidic synthesis reactions are heavily represented as the top contributing variables to variance.

Concerning the individuals graph, a similar pattern to the one analyzed previously can be observed, where there is a clear separation between formulations 1P and 2, in contrast to formulations 1R, 3R and 4. There is also a point in space that has many possible formulations overlapped and that corresponds to the WTs. All overlapping data points suggest a strong grouping of equal phenotypes as usual, and may be confirmed through HCA. The few solutions from formulations 1R and 3R are, in fact, equal to each other which means they can be treated as unique solutions. A noticeable difference comparing to model B is that formulation 4 has its own independent grouping. Nevertheless, these solutions demonstrate similar behaviour to model B solutions where most have a reaction that concerns reducing power (*R_TRANSH2*) with combinations of PPP reactions (*R_6PGDH*, *R_TALA1*, *R_R5P1*, *R_TKT1*, *R_G1D* and *R_GLUCK*) and glycolysis/gluconeogenesis reactions (*R_PGI*, *R_PGM*, *R_PFK* and *R_ENO*). Furthermore, regarding the remaining two formulations, there is less grouping and more solution variety and these are the solutions more closely related to the WTs.

From the full HCA dendrogram (in Appendix C) it is possible to corroborate the separation visualized on the score plot. Once more, it is harder to find an evident cut-off height that can be helpful to separate different clusters as there are plenty of options. Regardless, it is possible to isolate four main groups as seen in the full dendrogram colour labelling. Overall, the addition

of the phosphotransferase production did not add too much variation in the system as before. The grouping is very closely related to Model B results and there are not many new options or solutions from this metabolic model.

Lastly, the top ten contributing knockouts concerning the overall model C solutions is summarized below in Table 3.7.

Table 3.7: Model CMM_C top ten most targeted reactions for knockouts in the overall solutions set (#KO).

Reaction	**# KO**
R_PGM	48
R_TRANHS2	39
R_PYK	38
R_PGDH	27
R_6PGDH	22
R_PYC	19
R_PGI	18
R_MTHFT // R_SDH	16
R_G1D // R_G6P1D // R_GLUCK // R_R5PI // R_TALA2 // R_TKT1	10
R_ENO // R_PFK	8

These statistics are very similar to the previous ones (from model B) with just some smaller differences regarding knockout order. Nevertheless, it is interesting to note that *R_TRANSH2* is still on top of the list. Overall, practically all reactions belong to a pathway that metabolizes glucose to pyruvic acid, with the exception that *R_MTHFT* and *R_SDH* belong to the one carbon unit and TCA cycle family, respectively.

3.1.2.4 Model D

In Model D, plasmid and resistance marker productions are dependent on plasmid availability and, thus only one formulation was computed (formulation 1R). A major difference in the enumeration problem for this model is that it was allowed a MCS size of 8 knockouts as there were not many solutions for smaller sized MCSs. From the initial 29 different solutions obtained for this problem, only 27.6 % remained for further analysis after processing steps. Table 3.8 summarizes the number of knockouts in a solution and the number of solutions that have that size.

Table 3.8: Model CMM_D Summary of the suggested knockout number in a solution (MCS Size) and the amount of solutions that have that size (# MCS), corresponding to each formulation before (pre-) and after (post-) processing steps.

	Formulation	MCS Size	# MCS
Pre-Processing	1R	1	0 (0%)
		2	0 (0%)
		3	0 (0%)
		4	2 (7%)
		5	2 (7%)
		6	11 (37.9%)
		7	5 (17.2%)
		8	9 (31%)
Total			29
Post-Processing	1R	1	0 (0%)
		2	0 (0%)
		3	0 (0%)
		4	2 (25%)
		5	0 (0%)
		6	0 (0%)
		7	0 (0%)
		8	6 (75%)
Total			8

It is noticeable that solutions have a size that ranges from four to eight knockouts, and are more concentrated around the six to eight knockout size. After the filtration step, solutions with a four MCS size remained the same. At the same time, solutions with a size of 8 knockouts were reduced from 9 to 6 and the remainder were completely removed by processing. From all the models, this is the one with less pre- and post-processed solutions and it could be interesting to understand in detail why there are fewer possibilities.

Moreover, to better understand and visualize these differences and results, the PCAs and HCAs performed are shown in Figure 3.4.

The scree plot shows that only three principal components are required to explain 100% of the data variance, which is the lowest comparing to all previous models. Thus, concerning model D data, two principal components were chosen to be analyzed as they account for nearly the totality of variation (98.2 % cumulative explained variance percentage). This may have happened due to low amount of solutions in comparison to the other enumeration problems, as

well as these solutions being possibly closely related regarding their variables (reaction fluxes) which will translate in low variation.

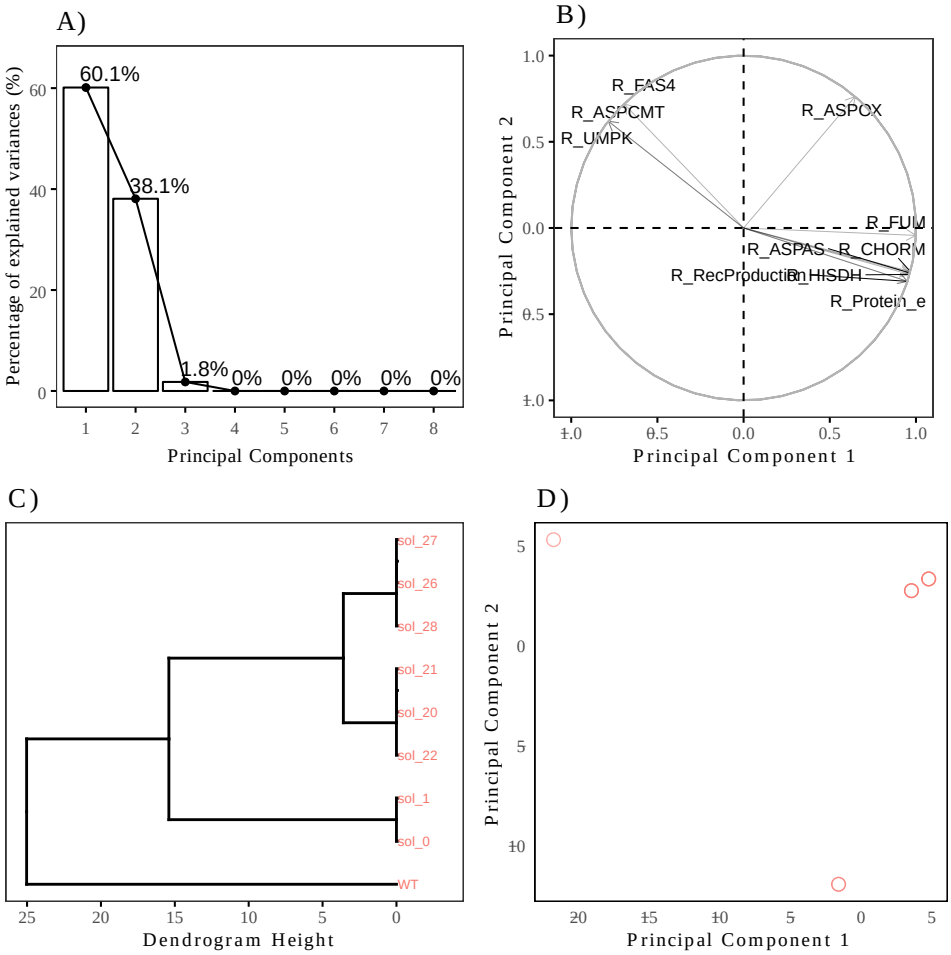

Figure 3.4: Model CMM_D Exploratory data analysis results: **A) Scree plot** percentage of explained variances (%) in each principal component (up to a total of ten PCs); **B) Correlation circle** correlation between the top 10 variables contributing to the PCs and the first and second principal components; **C) Dendrogram** hierarchical cluster analysis performed using single linkage method and euclidean distance metric. The sub-tree was obtained by a cut done at a dendrogram height equal to 15; **D) Individuals graph** data projection coordinates in the first two principal components: ○ CMM_D1R.

Concerning the correlation circle, from all the models, this is the one that shows more diversity. All top ten variables contribute equally to the components as previously stated and are separated in mainly three groups. The first group comprises *R_FAS4*, *R_ASPCMT* and *R_UMPK*, which are reactions that are related to nucleotide synthesis. Furthermore, a second group comprises the reactions *R_FUM*, *R_CHORM*, *R_ASPAS*, *R_HISDH*, *R_RecProduction* and *R_Protein_e*. Contrarily to the previous group, these reactions are all related to amino acid synthesis, mainly histidine, aspartate and aromatic families. These two groups share a 180 degree which means that they are negatively correlated. In solutions, when amino acids are being produced, it negatively influences nucleotide synthesis and vice-versa. This can be problematic since in this model recombinant protein production is completely dependent on plasmid availability. The last group is composed only by *R_ASPOX* which accounts for asparagine synthesis. This arrow is nearly orthogonal to the remaining vectors which implies that it does not have a strong

correlation to both groups. This result may be interesting as this reaction is tightly connected with *R_ASPAS* and, therefore requires further flux analysis.

Furthermore, from the score plot and dendrogram it is possible to observe three distinct clusters. In *sol_0* and *sol_1* group, the MCS size is equal to 4 knock outs, where 3 are contained in both solutions (*R_PEPCK*, *R_TRANSH2*, *R_PGI*) as the only difference lies on *R_GLUCK/R_G1D*. Moreover, for the middle and top clusters, the size increases to 8 knock-outs and these solutions share always six reactions in common (*R_6PGDH*, *R_PYK*, *R_PGM*, *R_MAL1*, *R_MAL2* and *R_G6P1D*) and the variation lies in PPP reactions *R_TALA2*, *R_R5PI* and *R_TKT1*. The solutions from this model can be easily clustered for further analysis which is an advantage in comparison to the previous models. However, eight knockout solutions may be too much to apply in a biological setting and thus, lower knock out reactions that present the same phenotype may be prioritized. Regardless, in the next Section 3.1.3 the core criteria behind which solutions are chosen for further detailed analysis will be described.

Lastly, since model D has fewer solutions, all the targeted reactions for knockouts concerning model D data are summarized below in Table 3.9.

Table 3.9: Model CMM_D top most targeted reactions for knockouts in the overall solutions set (#KO).

Reaction	**# KO**
R_6PGDH // R_G6P1D // R_MAL1 // R_MAL2 // R_PGM // R_PYK	6
R_PGI	5
R_PFK	3
R_PEPCK // R_R5PI // R_TALA2 // R_TKT1 // R_ TRANSH2	2
R_G1D // R_GLUCK	1

For model D targeted reactions, the pattern repeats itself where most reactions belong to glycolysis, ED or PPP. In this case, *R_TRANSH2* is not on top of list and the 'outliers' from this group are *R_MAL1* and *R_MAL2* that correspond to malic enzymes from the TCA cycle.

3.1.3 Detailed Network Analysis

In order to understand and find solutions that could be possible candidates for testing *in vivo*, a more detailed network analysis based on the pFBA fluxes was performed. For this, the first step was to identify two to three solutions that could be strong candidates based on what was explored in previous exploratory data analysis. The core idea is to compare the mutant flux pattern to the WT and try to understand where is the carbon source being allocated and what differences there are that seem relevant in a biological context. To find these solutions, a set of selection criteria was applied as follows: **(1)** biomass growth reaction with a positive non-zero flux; **(2)** priority to solutions with number of knockouts as low as possible; **(3)** avoid solutions

whose suggested knockouts are transport and exchange reactions; **(4)** priority to solutions that are highly represented in the MCS pool; and **(5)** if possible, allow some variability regarding suggested reactions pathways (for instance, a 2 KO solution with a reaction from fatty acid synthesis and one from glycolysis). These sets of criteria were all applied, with no specific order but rather in a way that it is possible to make a weighted and conscious decision.

That being said, the first MCS analyzed comprises the reactions *R_PGI* and *R_ENO* (MCS1). This solution appears in enumeration problems CMM_A1P, B1P, B2, C1P and C2. Moreover, this MCS follows most selection criteria and, in addition, is a good solution to compare to the previous work done by Pandey *et al.* (2018) as it suggests *pgi* knockout. These simulation results are presented in Figure 3.5, which displays a representation of the *E. coli* central carbon metabolism.

In this solution, by knocking-out these two reactions in the model, it was possible to produce plasmid with a 4.36 BPCY, while keeping the growth rate at 34.1% of the parental strain. In regard to the *pgi* knockout, since this reaction is a common node for different glucose catabolism pathways, its inactivation is particularly relevant for studying metabolic behaviour as carbon flux is redirected towards the PP pathway and/or the ED pathway. This flux rerouteing has a profound impact in redox balance where transhydrogenases have a critical role (Canonaco *et al.*, 2001). Moreover, concerning the *eno* knockout, this reaction is the penultimate step of glycolysis and catalyzes the reversible reaction between 2-phospho-D-glycerate and PEP. It is also a relevant reaction to study as it has an important role in gluconeogenesis. Regarding the latter knockout, there is a lack of experimental ^{13}C-fluxomics data, which can difficult the double knockout mutant flux distribution analysis (Long & Antoniewicz, 2014).

From the simulated flux distribution, it is possible to indicate that practically all glucose flux is redirected to the ED pathway and that there is not reallocation towards the oxidative PP pathway. In previous ^{13}C-MFA studies of a *pgi*-knockout strain it was experimentally determined that the PP pathway was the major route for glucose metabolism, providing a high NADPH source. Nevertheless, the ED pathway was also actively catalyzing a minor fraction of glucose in both wild-type and mutant strains (Hua *et al.*,2003; Fischer & Sauer, 2003). Although this single *pgi*-knockout MFA experimental results do not match the predicted pFBA flux distributions, it is important to take into consideration that our simulations concern a double knockout. Thus, the *eno*-knockout may present an important role in flux redirection. It is possible that, in our MCS simulation, a carbon flux allocation priority is shifted towards ED pathway as it is a more direct way to obtain T3P readily available to subsequently produce, for instance, serine family amino acids.

Moreover, in a study performed by Canonaco and Sauer it was shown that *pgi* inactivation led to a drastically reduction in maximum growth rate from 0.74 to 0.16 h^{-1}. In this mutant, an accumulation of NADPH due to an insufficient re-oxidation was also observed. The deficit observed in the growth rate was partly recovered by overexpressing the soluble transhydrogenase UdhA. Since this enzyme is responsible for converting NADPH into NADH, there is a probability that the growth recovery was due to the restored redox balance. In a cell, the redox balance is mainly described by the ratios between NAD^+/NADH and $NADP^+$/NADPH.

MCS1 - {R_PGI; R_ENO}

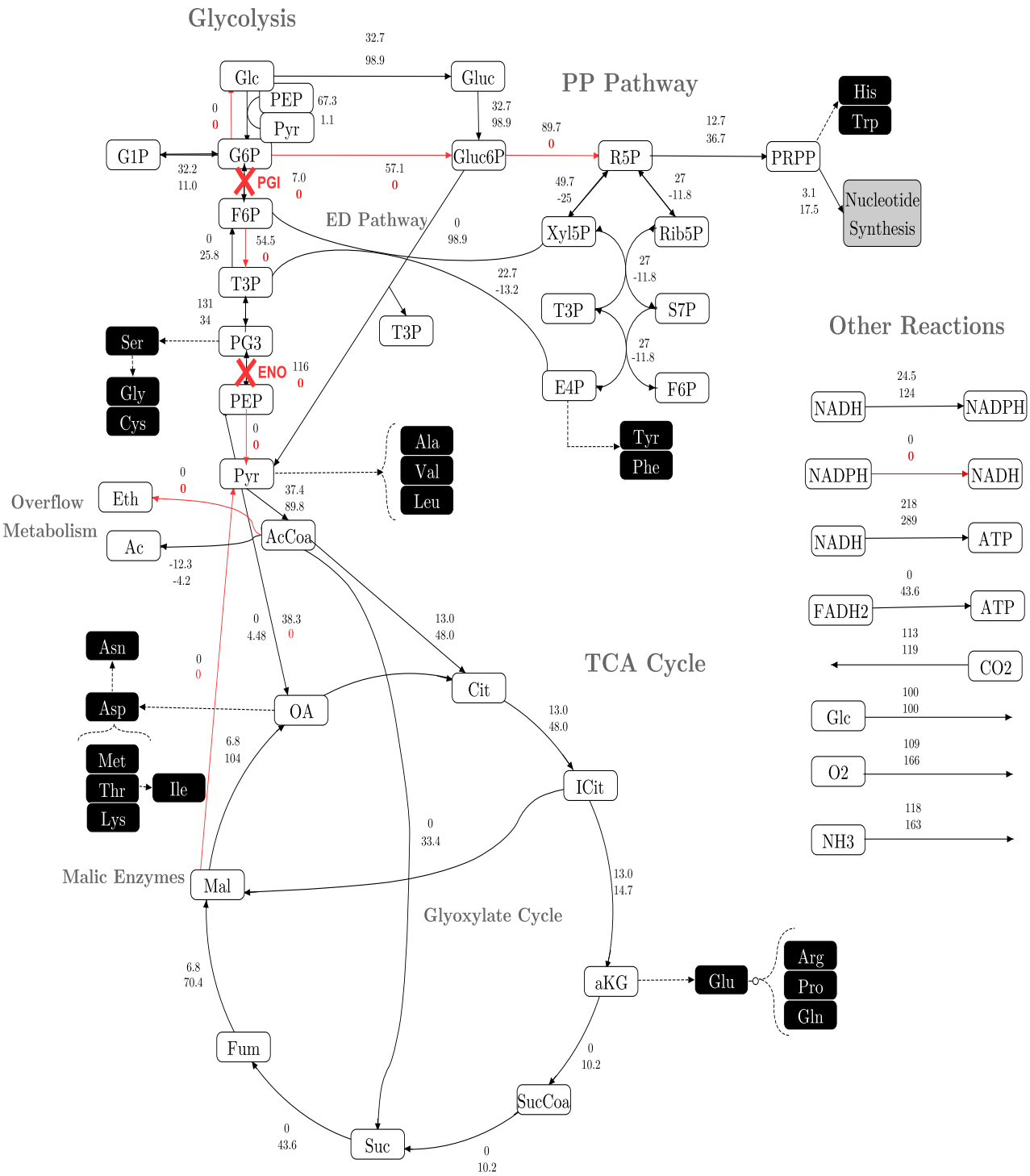

Figure 3.5: MCS1 metabolic flux distribution within central carbon metabolism of *E. coli* wild-type (top values) and Δ*pgi*Δ*eno* double knockout mutant (bottom values). Fluxes are given relative to the specific glucose consumption rate of 100 $mmol/g \cdot h$ and are expressed as the net fluxes. Knocked-out reactions are highlighted by a red cross and respective reaction name. Reactions from the mutant pFBA distributions that did not present flux were highlighted with red. Arrows indicate the directions of the proposed metabolic model (negative fluxes correspond to the inverse reaction). For abbreviations and detailed reactions, *vide* Appendix A.

These molecules participate in oxidation-reduction reactions and are specialized in carrying high-energy electrons and hydrogens, while transferring them to different sets of molecules. The main difference between these two molecules lies in NADH being mostly used in catabolic pathways and NADPH in anabolic pathways. Concerning catabolic reactions, NAD^+ serves as an oxidizing agent and is reduced to NADH whereas, in anabolic reactions, NADPH serves as a reducing agent and provides high-energy electron being reduced to $NADP^+$. The difference of a single phosphate group has no effect in both molecules redox properties; however, it helps enzymes distinguish these substrates. This is important so that both catabolic and anabolic pathways can be independently regulated, preventing futile metabolic cycles (Alberts *et al.*, 2002; Berg *et al.*, 2002). Considering Canonaco and Sauer experimental results, the fact that in our simulation the PP pathway is inactive can represent an advantage to the cell, since prevents excessive NADPH accumulation and a potential redox unbalance. However, the cell still requires a NADPH source to support anabolic metabolism.

When the PPP is inactive, NADPH production can potentially be achieved by three different routes in *E. coli* : **(1)** the NADPH dependent malic enzyme; **(2)** the membrane-bound transhydrogenase PntAB; and **(3)** the soluble transhydrogenase UdhA (Canonaco *et al.*, 2001). The first hypothesis is not feasible as our double-knockout pFBA flux distributions (in Figure 3.5) show that the reactions regarding malic enzymes are inactive (reaction *R_MAL1* and *R_MAL2*). This is supported by experimental evidence that demonstrates that in *pgi*-knockout strains there is no malic enzyme activity(Canonaco *et al.*, 2001). Additionally, the behavior observed in this single knockout is expectable to be seen in the double mutant. Concerning options 2 and 3, in our metabolic network, these re-oxidation mechanisms are represented as two distinct reactions (*R_TRANSH1* and *R_TRANSH2*). From our results, it is possible to see that the flux towards *R_TRANSH1*, which generates NADPH from NADH, is one of the highest. In fact, this transhydrogenase activation is our main NADPH source as it accounts for 78% of total NADPH pool. The remaining 22% are solely allocated from 5,10-methenyltetrahydrofolate (MeTHF) production reaction (*R_MTHFD*), since there is no carbon flux directed towards the oxidative branch of PP pathway.

Since in the flux distribution of the *pgi* and *eno* double knockout mutant, the NADPH availability is dependent on NADH pool, it is important to understand its source. In our simulation, NADH accumulation is mostly originated via TCA cycle (31.7 %) and via glycolytic pathway (30.5 %). Comparing with the WT simulation, an increment in the TCA cycle flux is observed which can explain NADH availability in the mutant. In particular, the flux in the conversion of malate into oxaloacetate is increased by 15-fold, providing a good NADH source. Contrarily to what was observed in the simulations with the WT strain, the glyoxylate shunt flux was activated in this double mutant. This is corroborated by some findings in a study performed by Usui et *al.* . The authors reported a sequential increment in the flux through the glyoxylate shunt as the phosphoglucose isomerase was successively down-expressed until it was completely knocked-out. It is known that in *E. coli*, the glyoxylate shunt is utilized mainly for the supply of oxaloacetate to the TCA cycle via malate by using isocitrate and acetyl-CoA (Kondrashov *et al.*, 2006). Thus, the activation of the glyoxylate shunt in the mutant strain increases malate availability, which in its turn is converted to oxaloacetate releasing NADH. That being said,

probably in this simulation, the glyoxylate shunt activation is essential to provide: **(1)** extra NADH to fulfill the NADPH requirements of the cell; and *(2)* oxaloacetate, that is an important precursor to a large family of amino acids, some of which are required in the nucleotide synthesis (such as L-aspartate).

This solution was generated in the model that only contemplates plasmid production, thus it is important to understand the flux allocation into nucleotide synthesis. The metabolite ribose 5-phosphate (R5P) of the PP pathway is the common building block in the *de novo* purine and pyrimidine synthesis pathways (Moffatt & Ashihara, 2003). In our simulation, this metabolite is generated by a reverse path through the non-oxidative PP pathway branch starting from the T3P generated in the ED pathway. From R5P, the flux is then directed towards PRPP, a common precursor to nucleotide synthesis (Moffatt & Ashihara, 2003). In the pFBA simulation results from the mutant, the flux increases in the previously described reactions with a consequent increment in nucleotide synthesis. Comparing with the WT flux values, there is an average 27-fold increase in the flux towards dNTPs synthesis reactions. In addition to nucleotide synthesis, energy expenditure concerning nucleotidic bonding needs to be taken into consideration (Equation 2.1). This means that, in our simulations, the flux of ATP must match this nucleotide synthesis increment to lead to a higher plasmid production. From the double mutant knockout pFBA results, it is possible to conclude that the TCA cycle operates predominantly for ATP generation by producing NADH that goes through oxidative phosphorylation. This is corroborated by the model reactions regarding oxidative phosphorylation (*R_ATPS1* and *R_ATPS2* that are NADH and $FADH_2$ dependent, respectively) accounting for approximately 88.1% of ATP generation flux. In particular, it is interesting to note that in the WT, $FADH_2$ production via TCA cycle is non-existent, whereas in the mutant it becomes an important energy source. Additionally, in the double knockout mutant, since the glycolytic pathway is mostly inactive, it provides only 11.4% of the energy source to the system.

Overall, this MCS is helpful in corroborating the findings by Pandey *et al.* even if the flux distribution does not fully match the experimental results. Nevertheless, it is necessary to take into consideration that our results are based on *pgi* and *eno* knockouts, instead of single knockout mutants. In spite of that, our double mutant did improve *in silico* plasmid production. However, it would be interesting to compare this simulation with experimental data from ^{13}C-MFA of single *eno*-knockout strains as well as double *pgi* and *eno* knockout strains to confirm, for instance, if the flux is preferably allocated towards ED pathway and how it impacts NADH/NADPH pool availability. In addition, contingent on the results from the single and/or double knockouts, it could be interesting to study the soluble transhydrogenase UdhA expression with the purpose to verify and corroborate its kinetic limitations in cell growth, plasmid and/or recombinant protein production.

Moreover, a second MCS was analyzed in detail that comprises reactions *R_PGM*, *R_G6P1D*, *R_PYK* and *R_PYC*. This solution appears in enumeration problems CMM_B1P, B2, C1P and C2. Moreover, this MCS matches most selection criteria and, from all the solutions in all models, is the one that presents the highest BPCY (regarding plasmid production). Additionally, the suggested knockouts show a reasonable variability regarding their role in the metabolism. These simulation results are presented in Figure 3.6.

MCS2 - {R_PGM; R_G6P1D; R_PYK; R_PYC}

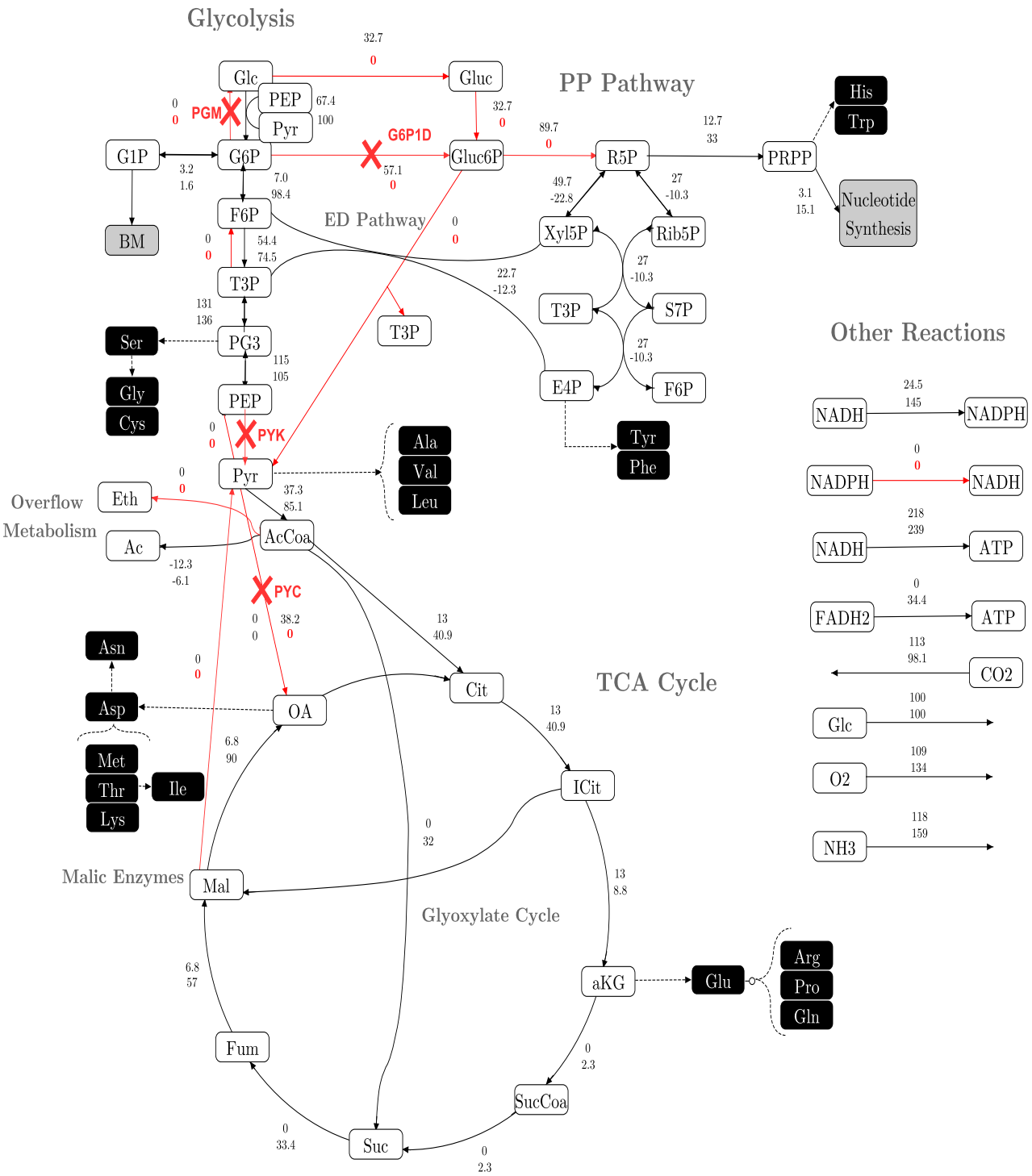

Figure 3.6: MCS2 metabolic flux distribution within central carbon metabolism of *E. coli* wild-type (top values) and $\Delta pgm/agp\Delta pykA/pykF\Delta zwf\Delta ppc$ quadruple knockout mutant (bottom values). Fluxes are given relative to the specific glucose consumption rate of 100 $mmol/g \cdot h$ and are expressed as the net fluxes. Knocked-out reactions are highlighted by a red cross and respective reaction name. Reactions from the mutant pFBA distributions that did not show flux were highlighted with red. Arrows indicate the directions of the proposed metabolic model (negative fluxes correspond to the inverse reaction). For abbreviations and detailed reactions, *vide* Appendix A.

In this solution, by knocking-out these four reactions in the model, it was possible to produce plasmid with a 5.23 BPCY, while keeping the growth rate at 49.8% of the WT strain. However, there was no flux going through the recombinant protein and resistance marker production reactions. In these models (B and C), the plasmid, resistance marker and recombinant protein synthesis share common precursors. Thus, by optimizing the flux through one of these products, the increasing required pool of precursors will limit the synthesis of the remaining reactions. For this reason, it is rather difficult to obtain solutions where production of 2 or 3 of these products take place at the same time.

The inactivated reactions from this MCS comprise two reactions from glycolysis and gluconeogenesis, namely *R_PGM* and *R_PYK* that are encoded by genes *pgm/agp* and *pykA/pykF*, respectively. In addition, *R_G6P1D* is the first step in the oxidative PP pathway, catalyzed by the enzyme glucose-6-phosphate 1-dehydrogenase that is encoded by gene *zwf*, and *R_PYC* that is an anaplerotic reaction whose catalytic enzyme is encoded by the *ppc* gene.

Accordingly to the flux distribution shown in Figure 3.6, the *pgm/agp* and *zwf* inactivation, led to a rewire of the flux towards glycolysis in comparison to the previous simulation results. Additionally, the oxidative PP Pathway and ED pathway had no flux. In ^{13}C-MFA studies of *zwf*-knockout strains, it was revealed that the disruption of glucose 6-phosphate dehydrogenase was counteracted by local rerouting via the glycolysis. Additionally, the authors have shown that the mutant strain synthesized the PPP-derived compounds independently from the oxidative branch by directing the carbon flow from glycolysis into the reversed non-oxidative PPP branch. Thus, it indicates that the glycolytic metabolites triose 3-phosphate and fructose-6-phosphate compensated the lack of E4P and R5P (Hua *et al.*, 2003; Nicolas *et al.*, 2007). Regarding *pgm/agp*-knockout, there is a lack of biological fluxomics data, which can difficult the interpretation of their role and behavior in our quadruple-knockout mutant.

Considering the data retrieved from the literature, the redirection of all carbon flux towards glycolysis, avoiding the oxidative PP pathway, can result in a NADPH shortage - an essential cofactor for anabolic metabolism (Nicolas *et al.*, 2007). Thus, the *zwf*-knockout strains must have a coping mechanism tightly related to transhydrogenases and NADP(+)-dependent enzymes. Zhao et *al.*, concluded that, as a response to this unbalance, the cell would use NADP(+)-dependent isocitrate dehydrogenase. This enzyme catalyzes the conversion of isocitrate into 2-oxoglutarate while producing the majority of NADPH. However, the action of this single enzyme is not sufficient to fulfil the NADPH cellular requirements, thus activating other enzymes such as transhydrogenases and NADP(+)-dependent malic enzyme is required. In our metabolic model, the reaction catalyzed by NADP(+)-dependent isocitrate dehydrogenase is not represented, which may explain some discrepancies observed between experimental data and simulation results. Similarly to what was observed in the MCS1 simulation, the source of NADPH is a consequence of the flux going through *R_TRANSH1*, that is responsible in converting NADH to NADPH. In fact, this transhydrogenase activation is responsible for approximately 82.5% of NADPH pool, while the remaining percentage is due to 5,10-methenyltetrahydrofolate production reaction (*R_MTHFD*). Additionally, it was observed that the reactions catalyzed by the malic enzymes have no flux. These enzymes catalyze NADH and NADPH production by converting malate into pyruvate. In our simulation results, there is a high pyruvate accumula-

tion, derived from the PTS system (*R_PTS*). This system is a distinct method used by bacteria for sugar (namely glucose) uptake where the source of energy comes from PEP. For each molecule of glucose that enters the cell, a molecule of pyruvate is produced. Thus, it is expected that it is not necessary to overproduce pyruvate by using the malic enzymes. Instead, in our simulation, malate is converted into oxaloacetate releasing NADH that is later converted into NADPH by transhydrogenases (*R_TRANSH1*). Nevertheless, despite our results not showing evidence of malic enzyme activity, experimental evidence demonstrates that *zwf*-knockout strains allocated 3% of their total carbon flux towards malic enzyme pathway (Zhao *et al.*, 2004).

Moreover, focusing on *pykA* and *pykF* mutants, a study by Fischer and Sauer shows that the metabolic bypass of pyruvate kinase knockout is done via PEP carboxylase (*ppc*) and malic enzyme as the mutants exhibit lower fractions of oxaloacetate originated through the TCA cycle and higher fractions of pyruvate originated from malate. In addition to the *R_PYC* anaplerotic reaction, it was demonstrated a depletion of *Pfk* (*R_PFK*) by the accumulated PEP. Even though these genes are knocked out, PEP can still be converted to pyruvate through PTS and, in the *pyk* mutants, the ATP level from aerobic respiration is not significantly affected *in vivo* (Zhu & Shimizu, 2005). Another important finding concerns excess NADPH produced by the *pyk*-knockouts mutants, which is similar to *pgi*-knockouts strains that tend to overproduce NADPH and convert the excess NADPH into NADH using transhydrogenases (Toya *et al.*, 2010). However, in both simulation and experimental data, the *zwf* mutant activates the reversed non-oxidative PP pathway and no flux passes through the oxidative portion, thereby avoiding NADPH accumulation.

Furthermore, it was previously mentioned that *pykA* and *pykF* mutants would increase PEP carboxylase (*R_PYC*) activity as it was a metabolic bypass for PEP metabolism. However, in our MCS solution, this enzyme is also inactivated so it is important to understand the effects on the metabolism. Evidences from a study done by Fong *et al.* reveal that, in *ppc* mutants, the normally repressed glyoxylate replaced the anaplerotic function of PEP carboxylase. Even though, in different studies regarding single knockouts of *zwf* and *pyk* mutants, the glyoxylate pathway is inactive and negligible, in the simulation results from MCS2 this reaction is activated. This may be due to PEP carboxylase anaplerotic role substitution, as well as additional NADH production that may be converted to NADPH via transhydrogenases. This mechanism is similar with what is previously observed in MCS1 solution, where NADH pool is originated from the TCA cycle, representing a flux increment between 3 and 13-fold in the reactions constituting this cycle when compared with the WT results. In particular, the flux of the conversion from malate to oxaloacetate suffered a 13-fold increment, due to glyoxylate shunt activation, which increased the malate pool. Additionally, oxaloacetate is an amino acids precursor and is tightly involved in nucleotide synthesis and consequent plasmid accumulation. Thus, this bypass is also important as it is the only source to indirectly produce oxaloacetate in our metabolic model.

Regarding nucleotide synthesis, it was observed a similar behavior to the previous results from MCS1. The flux starts from T3P and F6P and goes through the non-oxidative PP pathway until R5P is formed, which is then converted to PRPP (a common precursor to the nucleotide synthesis). Comparing to the WT pFBA results, it is possible to conclude that the flux towards dNTPS synthesis reactions is increased 22 to 23-fold. Considering Equation 2.1, to concomitantly

increase the production of plasmid, an increment in the ATP pool is also required. Similarly, from these results it is possible to conclude that the TCA cycle operates predominantly for ATP generation by producing NADH that goes through oxidative phosphorylation. This is corroborated by the model reactions regarding oxidative phosphorylation *R_ATPS1 and R_ATPS2* accounting for approximately 66.3 % of ATP generation. In particular, comparing the mutant to the WT, $FADH_2$ production plays a relevant role in energy source as its reaction net flux increases from 0.0 to 34.4. Nevertheless, it is important to note that, in comparison to MCS1, this simulation shows that one third of the ATP is also generated from glycolysis, as this pathway is not blocked.

In summary, in this quadruple-knockout the main goal seems to be to redirect flux as much as possible from G6P onwards and towards the glycolytic pathway. It is a solution that presents similar pFBA results to the previous one, however it has its differences. Namely, the carbon flux blockage towards the oxidative PP pathway (induced by the *zwf* deletion) that prevents NADPH overproduction, which is proven to be a limiting step towards growth and production in other mutants, such as single *pgi*-knockout strains. Additionally, these simulation results may be useful to explain some biological phenomena such as the glyoxylate shunt activation by analyzing the *ppc* and *pyk* mutants synergy. Overall, it would be interesting to possibly combine some of the findings from this MCS with the previous results, in order to construct a biological relevant and well supported MCS. In addition, having more biological data on more than single knockouts would be also useful to understand the metabolism and how the carbon flux is shifted.

Furthermore, a third and final MCS was analyzed in detail and comprises reactions *R_PEPCK*, *R_TRANSH2*, *R_PGI* and *R_GLUCK*. This solution appears in enumeration problems CMM_B1R, B3R, B4, C1R and C3. Moreover, this MCS matches most selection criteria. Contrarily to the previous MCS1 and MCS2, this solution has flux going through recombinant protein production reaction, instead of plasmid production. Additionally, the suggested knockouts show a reasonable variability regarding their role in the metabolism. These simulations results are presented in Figure 3.7 in an *E. coli* central carbon metabolism representation.

In this solution, by knocking-out these four reactions in the model, it was possible to produce IFNγ with a 1.30 BPCY, while keeping the growth rate at a 14.6% of the parental strain. However, there was no flux going through the plasmid and phosphotransferase production reactions. This is due to the fact that this solution was originated from a formulation problem that considers recombinant protein production optimization. Hence, it is rather difficult to obtain solutions where production of 2 or 3 of these products take place at the same time.

The inactivated reactions from this solution comprise a reaction from glycolysis *R_PGI* that is encoded by *pgi* gene; a reaction concerning an anaplerotic pathway (*R_PEPCK*) that is encoded by *pck* gene; a step in gluconate metabolism catalyzed by the enzyme gluconokinase (*R_GLUCK*) that is encoded by genes *idnK* or *gntK*; and a reaction regarding NADH regenerating through NADPH (*R_TRANSH2*), that is catalyzed by a transhydrogenase which is encoded by *pntAB* or *udhA* genes.

MCS3 - {R_PEPCK; R_TRANSH2; R_PGI; R_GLUCK}

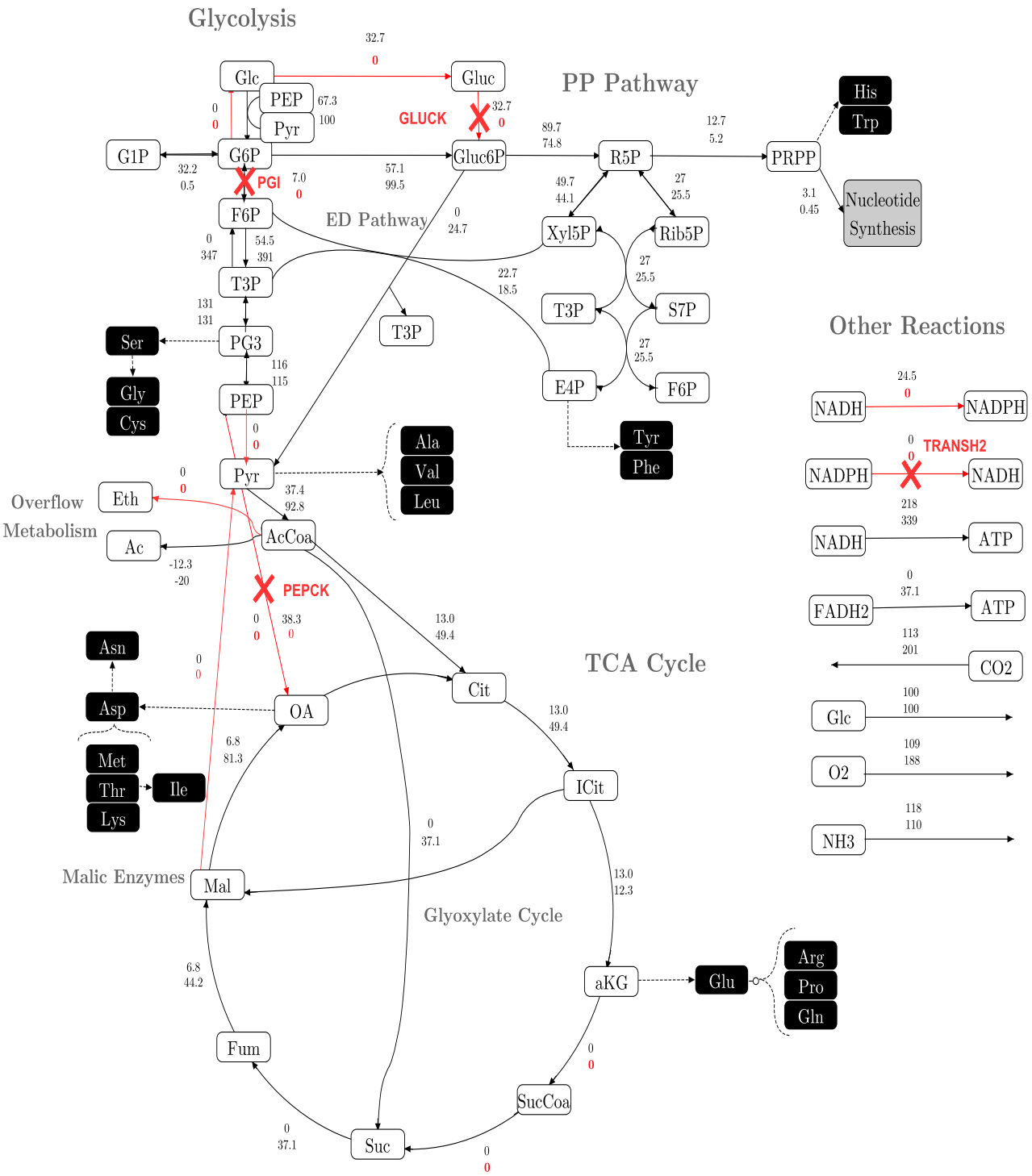

Figure 3.7: MCS3 metabolic flux distribution within central carbon metabolism of *E. coli* wild-type (top values) and $\Delta pck\Delta pgi\Delta pntAB/udhA\Delta idnK/gntK$ quadruple knockout mutant (bottom values). Fluxes are given relative to the specific glucose consumption rate of 100 $mmol/g \cdot h$ and are expressed as the net fluxes. Knocked-out reactions are highlighted by a red cross and respective reaction name. Reactions from the mutant pFBA distributions that did not show flux were highlighted with red. Arrows indicate the directions of the proposed metabolic model (negative fluxes correspond to the inverse reaction). For abbreviations and detailed reactions, *vide* Appendix A.

According to the flux distributions in Figure 3.7, the *pgi* and *idnK/gntK* inactivation led to a rewire of the carbon flux towards the ED and PP pathways. Considering these pathways, there was a 74.8% carbon allocation towards oxidative branch of PP pathway, while the remaining flux was redirected towards ED pathway. These predicted pFBA flux distributions are in accordance with previous ^{13}C-MFA studies of a *pgi*-knockout strain, where it was experimentally validated that the PP pathway was the major route for glucose metabolism after knocking-out *pgi* gene. In addition, these experimental studies proved that the ED pathway was also actively catalyzing a minor glucose fraction (Fischer & Sauer, 2003). Regarding *idnK/gntK*-knockout, there is a lack of biological fluxomics data, which can difficult the interpretation of its role in our quadruple-knockout mutant. Nevertheless, in our simulation results, it seems that this knockout mostly reinforces the carbon flux redirection towards PP and ED pathways.

Considering our quadruple-knockout mutant pFBA flux distributions, there is a high NADPH production due to a flux allocation towards the PP pathway. Nearly 97.8% of NADPH is produced in this pathway, while the remaining 2.2% are from 5,10-methenyltetrahydrofolate (MeTHF) production reaction (*R_MTHFD*). NADPH is an important cofactor for anabolic reactions. To increase recombinant protein production, a concomitant increment in amino acids pool is also required. Consequently, to produce these amino acids, a higher NADPH pool is necessary. In our simulations, the conversion of NADPH into NADH, catalyzed by reaction *R_TRANSH2*, is knocked out. This way, all NADPH generated through the PP pathway can be allocated towards biosynthetic pathways (such as amino acids precursors synthesis). However, experimental data retrieved from literature shows that carbon flux redirection to PP pathway leads to an accumulation of NADPH due to an insufficient re-oxidation (Canonaco *et al.*, 2001). This accumulation led to a reduction in growth rate that was later partly recovered by overexpressing the soluble transhydrogenase UdhA. This enzyme is responsible for converting NADPH into NADH, hence there is a probability that growth recovery was due to the restored redox balance, as previously described in MCS1. In our metabolic model, this mechanism is inactivated (*R_TRANSH2*) and thus, our model is unable to re-oxidize NADPH through this reaction that is catalyzed by transhydrogenase. Therefore, our simulation results may not correspond to a feasible biological state. Since *pgi*-knockouts were experimentally proven to accumulate NADPH, it is probable that a double *pntAB/udhA* and *pgi*-knockout is not able to strive in growth. Nevertheless, according to the amino acid synthesis requirements (Appendix B), and since we want to improve plasmid and recombinant protein production, it is understandable that the suggested knockouts try to increase cofactors pool such as NADPH. Hence, this solution could be a suggestion to test *in vivo*, as accumulated NADPH could be induced and redirected towards biosynthetic pathways.

Furthermore, in our simulation results, the flux is then directed towards the bottom half part of glycolysis and towards the TCA cycle. It is important to note in Figure 3.7 that the reaction interconverting F6P and T3P shows a higher amount of net flux in comparison to the remaining reactions. This is due to a futile cycle in this interconversion. It can be considered that the real flux is given by the subtraction of fluxes and, thus this reaction is preferably going in the forward direction. Entering an interrupted TCA cycle, in comparison to the WT, there is an increment on flux towards alpha-ketoglutarate formation (aKG) that is completely rewired towards glu-

tamic acid amino acids family production with no further conversion into SucCoa. Additionally, this increment towards aKG is accompanied by glyoxylate shunt activation. This activation leads to a flux re-allocation towards malate and succinate leading to a higher accumulation of oxaloacetate that is a precursor to aspartic acid amino acids family. Hence, both of these mechanisms are essential to accumulate important biosynthetic precursors towards recombinant protein production. This is also corroborated by the PEP carboxykinase knockout (*R_PEPCK*) as it prevents oxaloacetate decarboxylation into PEP, increasing even more its availability to the synthesis of these precursors. The results from an experimental study performed by Yang *et al.,* (2003), proved that *pck*-inactivation led to glyoxylate shunt activation to participate in anaplerosis and replenish the TCA cycle. Hence, the experimental results from the literature support the flux distributions obtained in our simulations.

Regarding IFNγ synthesis, most of the fluxes directed towards amino acids synthesis are increased when comparing with the WT simulation results. Comparing with the WT flux values, there is an average 1.5 to 2-fold increase through many amino acid synthesis reactions. These reactions are related to amino acids whose demand differs a lot from biomass to recombinant protein production. Some examples of these amino acids are lysine, serine, phenylalanine, histidine and leucine. In addition to amino acids synthesis, energy expenditure concerning peptidic bonding needs to be considered. This means that, in our simulation, the flux of ATP must follow this amino acid synthesis increment to effectively lead to a higher recombinant protein production. From the quadruple mutant knockout simulation results, it is possible to conclude that ATP generation is predominantly provided by oxidative phosphorylation as it accounts for 74.2% of the energy source (reactions *R_ATPS1* and *R_ATPS2* that are NADH and $FADH_2$ dependent, respectively). The NADH required for aerobic respiration is mostly provided by glycolysis (64.5%), while some is produced from the TCA cycle (27.0%). Moreover, $FADH_2$ production is exclusively a result of succinate dehydrogenase activity in the TCA cycle.

Overall, in this quadruple-knockout the flux is redirected towards PP pathway with consequent NADPH accumulation due to transhydrogenase inactivation. Additionally, in comparison to the WT simulation results, the higher flux through TCA cycle increases the amino acids synthesis precursors such as oxaloacetate and alpha-ketoglutarate. From experimental data in the literature, the *in vivo* application of these results probably will affect the maximum growth but enhance plasmid and recombinant protein production.

3.2 Genome-scale Model

The GSM used to generate results was iJO136, whose reconstruction was done by Orth *et al.*, 2011. The updated version of this model comprises 1367 associated genes, 2585 metabolic reactions and 1805 metabolites. In comparison to the CMM model, these metabolites and reactions can be compartmentalized, which is translated in another level of complexity and new type of reactions, such as transport reactions.

3.2.1 Data Processing

All data were attempted to be generated for each formulation problem as previously described in Chapter 2. However, formulations 2 and 4 will not be included in the GSM results as there were not enough computational resources to generate them. A maximum knockout size of 6 was allowed for each problem and all the solutions were stored as sets of strings encoding reactions. Equal to the CMM data generation, for each solution a pFBA flux distribution was computed and stored in a matrix.

Before analysing the data, a pre-processing step was performed to help reduce the number of solutions. In this case, a less broaden solution removal criterion was applied as the computational relevance was taken more into consideration than the biological one. Primarily, all solutions that were found and that were incompatible with the previously defined cellular constraints were automatically removed. In conjunction with this, solutions whose production occurred in a mandatory way coupled to 99 % of biomass were kept. Essentially, it was verified if all cellular constraints were followed and if the minimum product flux, with biomass growth fixed at 99% of its maximum value, was greater than zero.

For each model and formulation these filtration steps were applied. After processing, each formulation data corresponding to a model was concatenated and analysed simultaneously. Interestingly, it is important to note that model A (regarding plasmid production only) and model D (a more complex model correlating all entities) did not have any solutions that remained after this processing step and, thus will not be mentioned in the following and further data analysis. As this genome-scale network comprises more reactions and alternative pathways that may result in different ways to metabolize glucose, possible explanations lie on flux allocation. For instance, a possible hypothesis is that the GSM has a high number of pathway alternatives to nucleotide metabolism and is not able to make it essential in order to produce the plasmid, unless a considerable amount of knockouts are introduced. Another possibility relies on energy and the possible fact that the cell is not able to produce enough ATP to be used in the nucleotide synthesis pathway.

3.2.2 Exploratory Data Analysis

3.2.2.1 Model B

For model B possible enumeration problems, all formulations presented results. However, the solutions regarding plasmid production formulation (1P) did not survive the filtration step. On one hand, regarding 1R formulation, from an initial 1503 solutions, 195 remained for further analysis which corresponds approximately to 13% of the initial pool. On the other hand, for 3R formulation, from an initial higher solution number of 34 739, only 17 remained, corresponding to 0.049% of the initial pool, which makes it an extremely diluted solution pool comparing to the previous formulation. Overall, accounting for all formulations, from a total of 37501 initial solutions, only 0.57% (212) made it through the filtration process, which may be beneficial for further analysis. Table 3.10 summarizes the total number of solutions in each formulation for all enumeration problems performed on model B data.

Table 3.10: Model GSM_B Summary of the total number of solutions gathered (#Solutions), corresponding to each formulation before (pre-) and after (post-) processing steps.

	Formulation	**#Solutions**
Pre-Processing	1P	1 259
Post-Processing		0
Pre-Processing	1R	1 503
Post-Processing		195
Pre-Processing	3R	34 739
Post-Processing		17

From the post-processed set of 212 solutions there were a total of 15, 18 and 179 MCSs with a size of 4,5 and 6 suggested knockouts, respectively. Moreover, PCA and HCA were performed and the computed results for this set of solutions are shown below in Figure 3.8.

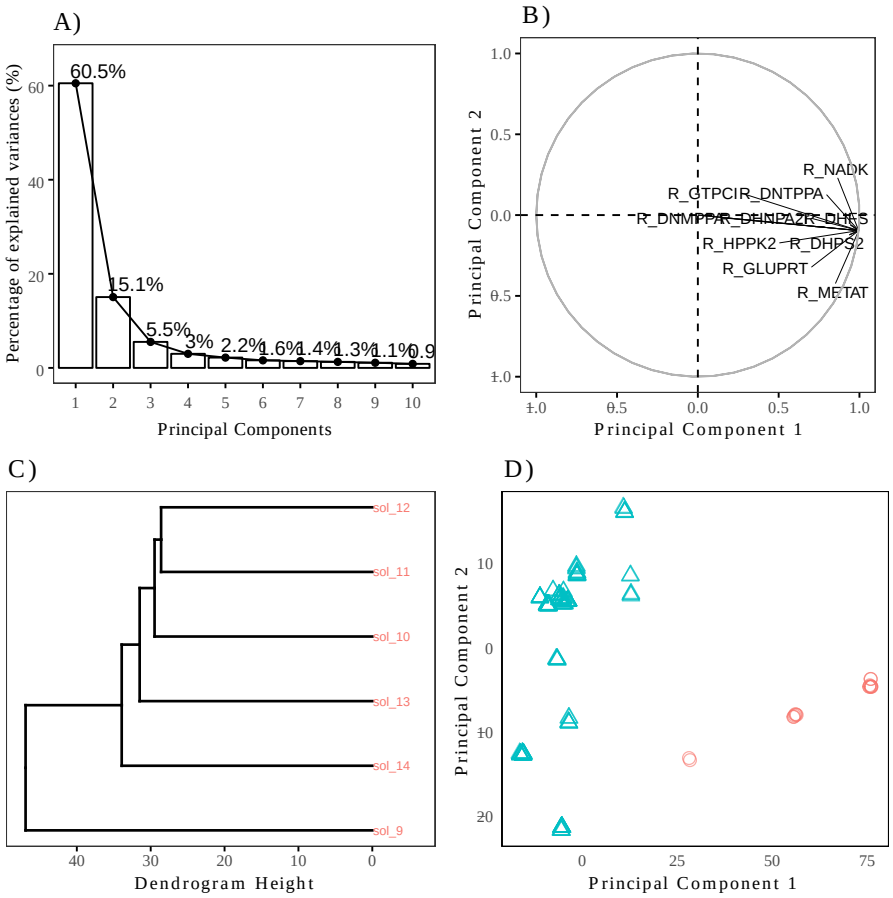

Figure 3.8: Model GSM_B Exploratory data analysis results: **A) Scree plot** percentage of explained variances (%) in each principal component (up to a total of ten PCs); **B) Correlation circle** correlation between the top 10 variables contributing to the PCs and the first and second principal components; **C) Dendrogram** hierarchical cluster analysis performed using single linkage method and euclidean distance metric. The sub-tree was obtained by a cut done at a dendrogram height equal to 50; **D) Individuals graph** data projection coordinates in the first two principal components: ○ GSM_B3R △ GSM_B1R .

From the scree plot is it possible to conclude that ten principal components are required to explain 92.6 % variance in these data, which is a lower percentage and higher dimensions required comparing to all previous CMM models. Nevertheless, if two principal components are chosen to represent the data, these account for 75.6 % of cumulative explained variance which is a reasonable amount of variation in two dimensions and is comparable to previous models. Therefore, concerning model B data, this was the amount of principal components chosen for further PCA analysis.

Moving to the correlation circle, it is possible to visualize that all top ten contributing variables share the same amount of contribution as their arrows share equal length. All ten variables are in the positive side of the first principal component and negative side of principal component 2, and demonstrate a strong positive correlation between each other. Moreover, seven out of ten variables correspond to reactions that belong to folate biosynthesis pathway (*R_DHFS*, *R_DHPS2*, *R_DNTPPA*, *R_DHNPA2*, *R_HPPK2*, *R_GTPCI* and *R_DNMPPA*) at different stages. Most of these reactions intervene in the synthesis of tetrahydrofolate, which is a cofactor in many reactions regarding, especially, *de novo* synthesis of purine and pyrimidine nucleotides and some amino acids interconversion, such as serine to glycine. The remainder reactions concern a NAD^+ kinase (*R_NADK*), which accounts for NAD^+ phosphorylation into $NADP^+$, which is an essential coenzyme that is reduced to NADPH primarily by the PP pathway to provide reducing power in biosynthetic processes such as fatty acid biosynthesis and nucleotide synthesis; a reaction regarding the conversion of PRPP into 5-phosphoribosyl-1-amine (PRA) using the ammonia group from the glutamine chain (*R_GLUPRT*) that is the committing step in *de novo* purine nucleotide synthesis; and S-adenosylmethionine synthesis reaction (*R_METAT*) that is an important methyl and propylamino donor in polyamine biosynthesis. Overall, the top contributing variables are concerning nucleotidic synthesis as previously described for the correlation plots from CMM models, with the difference that in this case, most of the reactions intervene in reducing power balance and folate synthesis intermediates.

Analyzing the individuals graph it is possible to see a separation between solutions from formulations 1R and 3R, meaning that there may be a different phenotype between these two formulations. It is also possible to note that GSM_B3R solutions are more clustered, whereas the remaining are more dispersed and show less of a pattern. This is also possible to corroborate from the full HCA dendrogram (in Appendix C), where it is possible to conclude that there is a high dissimilarity between solutions as there is a very high possible number of single clusters. Overall, in comparison to CMM, GSM has a higher number of reactions, thus there is additional levels of variability and more metabolism options which result in a high number of different solutions and phenotypes. This makes it harder to find a pattern and group solutions together, which was expected for the genome-scale model. Nevertheless, it is also important to note that comparing both models CMM_B and GSM_B, the suggested reactions for deletion in the MCSs (Table 3.11) are completely different from each other, which may be a suggestion that there are important alternative pathways that are not represented in CMM.

For model B it is possible to note that half of the most targeted reactions concern exchange or transport reactions (identified by the *tex* or *tpp* in the end) and that *R_Htex* is present in all 212 solutions. The remaining concern pyruvate reactions *R_PFL* and *R_PDH*, that regulate

the aerobic and anaerobic glucose metabolism, respectively, by converting pyruvate to acetyl-CoA. Additionally, there are malate related reactions *R_MDH*, *R_MOX* that correspond to the malate transformation into oxaloacetate by different means and with different cofactors to accept electrons and protons; and *R_SPODM* that is and important oxidative stress mechanism reaction.

Table 3.11: Model GSM_B top ten most targeted reactions for knockouts in the overall solutions set (#KO).

Reaction	**# KO**
R_Htex	212
R_CO2tpp	158
R_PFL	126
R_MOX	63
R_SPODM	63
R_FE2tex	62
R_FE3tex	62
R_FEROpp	61
R_MDH	57
R_PDH	27

3.2.2.2 Model C

For model C possible enumeration problems, some formulations did not present any results and only formulation 1R prevailed with solutions after the processing step. For the latter, from an initial set of 11 917 solutions, only 56 remained for further analysis after the filtration step which corresponds approximately to a 99.5% reduction. As a whole, comprising all formulations, from a total of 83 453 solutions, only a small percentage of 0.067% made it through the filtration process, making it a very diluted solution pool, which may be beneficial for further analysis. Table 3.12 summarizes the total number of solutions in each formulation for all enumeration problems performed on model C that had results.

From the post-processed set of solutions it is also important to note that, all were MCSs with a size equal to 6. Unfortunately, there were not any lower sized solutions. Moreover, a PCA and HCA were performed and the computed results for this set of solutions are show in Figure 3.9.

Table 3.12: Model GSM_C Summary of the total number of solutions gathered (#Solutions), corresponding to each formulation before (pre-) and after (post-) processing steps.

	Formulation	#Solutions
Pre-Processing	1P	35 768
Post-Processing		0
Pre-Processing	1R	11 917
Post-Processing		56
Pre-Processing	3R	35 768
Post-Processing		0

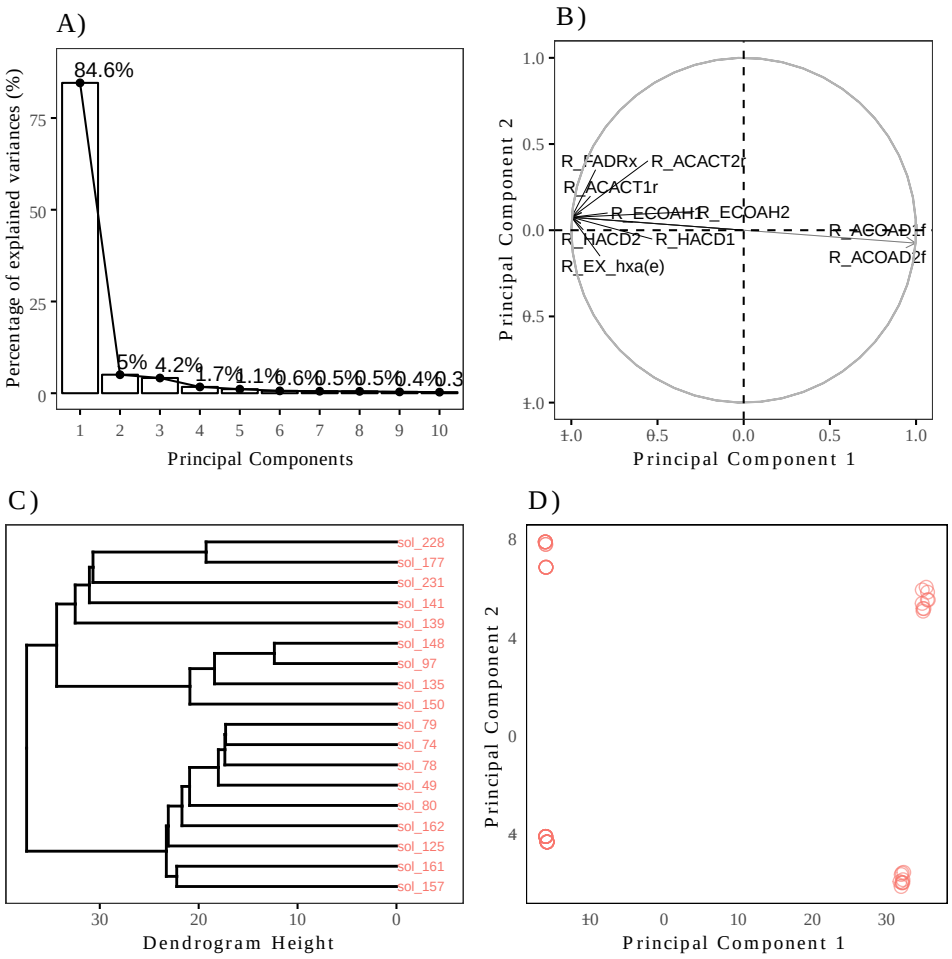

Figure 3.9: Model GSM_C Exploratory data analysis results: **A) Scree plot** percentage of explained variances (%) in each principal component (up to a total of ten PCs); **B) Correlation circle** correlation between the top 10 variables contributing to the PCs and the first and second principal components; **C) Dendrogram** hierarchical cluster analysis performed using single linkage method and euclidean distance metric. The sub-tree was obtained by a cut done at a dendrogram height equal to 50; **D) Individuals graph** data projection coordinates in the first two principal components: ○ GSM_C1R.

From the scree plot is it possible to conclude that five principal components are required to explain 96.6 % variance in these data, which is comparable to previous results. Interestingly, a single component accounts for 84.6 % of the explained variance which means that there is a very high degree of correlation between variables, or between at least two variables while the others show a much smaller dispersion. Geometrically, this translates in samples laying on a line in the space defined by the reactions. Concerning this model data analysis, two principals components were chosen to present the data as they account for 89.6 % of cumulative explained variance which is a reasonable amount.

As far as the correlation circle is concerned, all top tens variables equally contribute to the components as they show equivalent arrow length. Most of these reactions are in the negative side of principal component 1 and positive side of component 2, where only two variables show the opposite behaviour. These two groups of variables present a 180 degree angle, meaning that they are strongly negatively correlated. In comparison to the previous model, it is interesting to note that, with the addition of the resistance marker production, folate synthesis related reactions are no longer part of the top contributing variables. Instead, most reactions concern fatty acid metabolism, more specifically beta-oxidation, which is a process by which fatty acid molecules are broken down to generate acetyl-CoA that will enter the TCA cycle and produce energy down the line. These reactions concern mainly acyl-CoA dehydrogenation and hydration reactions that yield shortened fatty acids each cycle until acetyl-CoA is formed. Additionally, there is a reaction regarding FAD reductase (*R_FADRx*) that catalyses the conversion between $FADH_2$ and FAD, which is extremely important in restoring FAD pool, that is a necessary cofactor for acyl-CoA dehydrogenation. Interestingly, even though these reactions are all positively correlated to fatty acid synthesis, there is the *R_ACOAD1f* and *R_ACOAD2f* group that is shown to be negatively correlated and could be an interesting target for further analysis. Overall, the top contributing variables concern fatty acid metabolism reactions which can be a new perspective as these variables did not show up in previous results.

Furthermore, from the score plot it is possible to identify four different groups inside the same formulation. These solutions that are grouped together can be described by a unique linear combination of the reaction variables and demonstrate some symmetry between each other. This symmetry means that the data may be symmetric around its center. These overlapping points suggest a possible equal phenotype among them as they share in common many reactions, thus its similarity. For instance, reaction *R_Htex* is common to all solutions. Moreover, most of the reactions that were suggested for removal are exchange and transportation reactions (*R_O2tex*, *R_H2Stex*, *R_Htex*, *R_ETOHtex*, among others). Overall, there are four different patterns that could be analysed are shaped by combinations of reactions regarding mainly metabolite exchange and transport as well as some central carbon metabolism reactions, such as *R_MDH* (malate dehydrogenase), *R_MOX* (malate oxidase) and *R_PFL* (pyruvate formate lyase).

From the full HCA dendrogram (in Appendix C) it is possible to corroborate, from the score plot visualization, that there is separation in data points. A four cluster separation is not as evident but, at least a minimum of two or three main clusters may be delimited. It is interesting to note that the bottom half of the dendrogram shows larger, lesser and defined clusters, whereas the top half demonstrates higher entropy and more dissimilarity between solutions resulting in

less defined and higher amount of clusters.

Lastly, the top ten contributing knockouts concerning model C solutions are summarized below in Table 3.13.

Table 3.13: Model GSM_C top ten most targeted reactions for knockouts in the overall solutions set (#KO).

Reaction	**# KO**
R_Htex	56
R_O2tex	38
R_MDH	24
R_MOX	24
R_H2St1pp	19
R_H2Stex	19
R_PFL	18
R_ACALD	14
R_ACALDtex	12
R_ACALDtpp	12

For model C most targeted reactions it is possible to note that, as previously seen, half of them concern exchange or transport reactions and that *R_Htex* is present in all 56 solutions. The remaining concern malate and pyruvate metabolism reactions as described in model B. The only difference lies on this model having *R_ACALD* as a top ten target, which corresponds to acetaldehyde dehydrogenase (an important enzyme in alcohol metabolism). Overall, the GSM models rely and suggest much more knocking out exchange and transport reactions of metabolites that are important to the cell growth and metabolism. With that in mind, for a further detailed analysis, it would definitely need to be considered which solutions could be biologically relevant and not just part of the model and how mathematically formulated the model is. For instance, it would be interesting to remove solutions that account transport reactions in the data processing step as these are often essential and difficult to genetically manipulate in *in vivo* experiments.

Chapter 4

Conclusions

The work developed in this thesis was set out with the aim of applying a minimal cut set enumeration algorithm to find solutions for optimal and efficient plasmid and/or recombinant protein production (IFNγ in our case study). To accomplish this, a central carbon metabolism and genome-scale *E. coli* K-12 metabolic networks were used to perform simulations. To these models, a set of different ways to produce these compounds were added. In addition, different problem configurations were performed and, in the end, all results were concatenated and analyzed.

Concerning the central metabolism model results, from the exploratory data analysis, it was possible to observe a pattern regarding different formulations in most data for each model. Additionally, it was also possible to cluster some of the solutions that presented different knockouts but similar phenotypes, thus reducing the solutions pool size. From this analysis and a previously defined criteria, three solutions that were well represented were chosen for a further detailed analysis.

From the three examples solutions highlighted in the detailed network analysis section, a clear distinction between carbon flux allocations could be made. The choice of the first solution, MCS1, had as main goal to corroborate the findings from Pandey *et al.* (2018) that *E. coli pgi* mutant increased plasmid and recombinant protein production efficiency. This solution was a good candidate as it had a *pgi* knockout and suggested only one additional reaction for deletion (*eno* knockout). Even though the pFBA flux distribution did not fully match the findings by Pandey *et al.*, possibly due to the *eno* knockout effect in our double mutant, it was helpful to corroborate that plasmid production efficiency increased. Moreover, regarding MCS2 and MCS3, the main objective in the analysis was to identify a possible new knockout or set of knockout strategies that could lead to optimal production and seem biologically relevant and feasible. Accounting for all information collected in the pFBA flux distributions and experimental single-knockout studies some considerations can be made. *E. coli pgi* knockouts are proven to rewire carbon flux towards PP pathway which leads to a higher NADPH production (an important cofactor in anabolism). By knocking out the transhydrogenase activity, the interconversion between NADPH and NADH becomes blocked, resulting in NADPH accumulation. This metabolite pool can then be used to produce the necessary precursors for plasmid and recombinant protein production in higher quantities. Thus, a possible knockout to test *in vivo*

is transhydrogenases *udhA* or *ptnAB* genes. Even though, transhydrogenase inactivation was proven to affect cell growth, the NADPH accumulation may be beneficial to produce higher plasmid and recombinant protein yields. Moreover, another knockout that, from the detailed analysis, could be beneficial towards plasmid and protein production is PEP carboxykinase gene *pck* knockout. This may lead to a glyoxylate shunt activation and oxaloacetate accumulation (an important amino acid precursor) without compromising too much on maximum biomass growth. Overall, the genes *pgi*, *pck* and *udhA/ptnAB* seem promising to increase *in vivo* plasmid and/or recombinant protein production.

Regarding methodology, it can be concluded that from all model configurations, model A has more results with the lowest number of KOs. Nevertheless, models B and C also showed a good number of feasible KO suggestions. However, model D was the worst performing configuration when concerning to numbers of solutions and KOs. This may be due to the inability to properly balance the ratio between plasmid, recombinant protein and resistance marker production. Hence, from all configurations, having a single reaction accounting for plasmid and/or plasmid production seem to perform the best, depending on the production objective.

Concerning the genome-scale model, from the exploratory data analysis, it was possible to see that the suggested knockouts are completely different from the central carbon model ones. Additionally, most of the solutions were concerning exchange and transportation reactions. In a genome-scale model, there are more alternative pathways to where glucose focus may be redirected. When trying to force plasmid or recombinant production, there are more possible pathways that should be inactivated. As such, exchange and transportation reactions are a way to easily shut down these pathways to disrupt the balance in the system. However, these solutions are hardly applicable *in vivo* and thus, a further analysis would be helpful in understanding as to why these reactions were suggested to be knocked out.

4.1 Future Work

This work lacks a more detailed analysis regarding the genome-scale model, which could bring new insights not only to corroborate findings in the small model solutions, but also new sets of possible feasible knockouts. The production of other biotechonologically relevant products could also be a target of study, not only to draw conclusions regarding their flux allocation mechanisms but to explore new possible solutions. Finally, this work can be helpful in choosing a knockout or set of knockouts and, thus experimental validation and testing of this solutions would complement this work.

Bibliography

Alberts, B., Johnson, A., and Lewis, J. (2002). *Molecular Biology of the Cell. 4th edition. New York: Garland Science.*

Almquist, J., Cvijovic, M., Hatzimanikatis, V., and Nielsen, J. (2014). Kinetic models in industrial biotechnology - Improving cell factory performance. *Metabolic Engineering*, 24:38–60.

Archer, C. T., Kim, J. F., Jeong, H., Park, J. H., Vickers, C. E., Lee, S. Y., and Nielsen, L. K. (2011). The genome sequence of *E. coli* W (ATCC 9637): comparative genome analysis and an improved genome-scale reconstruction of *E. coli* . *BMC Genomics*, 12:9.

Baumler, D. J., Peplinski, R. G., Reed, J. L., Glasner, J. D., and Perna, N. T. (2011). The evolution of metabolic networks of *E. coli*. *BMC Systems Biology*, 5(1):182.

Becker, S. A., Feist, A. M., Mo, M. L., Hannum, G., Palsson, B. Ø., and Herrgard, M. J. (2007). Quantitative prediction of cellular metabolism with constraint-based models : the COBRA Toolbox. *Nature Protocols*, 2(3):727–738.

Berg, J., Tymockzo, J., and Stryer, L. (2002). *Biochemistry. 5th edition. New York: W H Freeman.*

Canonaco, F., Hess, T. A., Heri, S., Wang, T., Szyperski, T., and Y, U. S. (2001). Metabolic flux response to phosphoglucose isomerase knock-out in *Escherichia coli* and impact of overexpression of the soluble transhydrogenase UdhA. *FEMS Microbiology Letters*, 204:247–252.

Castrillo, J. I., Pir, P., and Oliver, S. G. (2013). *Yeast Systems Biology : Towards a Systems Understanding of Regulation of Eukaryotic Networks in Complex Diseases and Biotechnology.* Elsevier Inc., Cambridge.

Chalancon, G., Kruse, K., and Babu, M. M. (2013). Metabolic networks , structure and dynamics. In *Encyclopedia of Systems Biology*, volume 44, pages 1263–1267. Springer International Publishing, New York City.

Clark, D. P. and Pazdernik, N. J. (2015). Recombinant Proteins. In *Biotechnology : Applying the Genetic Revolution*, pages 335 – 363.

Clark, S. T. and Verwoerd, W. S. (2012). Minimal Cut Sets and the Use of Failure Modes in Metabolic Networks. *Metabolites*, 2:567–595.

Edwards, J. S. and Palsson, B. O. (2000). The *Escherichia coli* MG1655 *in silico* metabolic genotype : Its definition , characteristics , and capabilities. *PNAS*, 97(10):5528–5533.

Emmerling, M., Dauner, M., Ponti, A., Fiaux, J., Hochuli, M., Szyperski, T., Wu, K., Bailey, J. E., and Sauer, U. (2002). Metabolic Flux Responses to Pyruvate Kinase Knockout in *Escherichia coli*. *Journal of Bacteriology*, 184(1):152–164.

Estrada, F. L. (2010). *Interval and Possibilistic Methods for Constraint-Based Metabolic Models.* PhD thesis, Universidad Politécnica de Valencia, Spain.

Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., Broadbelt, L. J., Hatzimanikatis, V., and Palsson, B. Ø. (2007). A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Molecular Systems Biology*, 3(121):1–18.

Feist, Adam M., Herrgard, Markus J., Thiele, Ines, Reed, Jennie L. and Palsson, Bernhard Ø. (2009). Available predictive genome-scale metabolic network reconstructions. `http://systemsbiology.ucsd.edu/InSilicoOrganisms/OtherOrganisms`.

Figueiredo, L. F. D., Podhorski, A., Rubio, A., Kaleta, C., Beasley, J. E., Schuster, S., Planes, F. J., and Biology, C. (2009). Computing the shortest elementary flux modes in genome-scale metabolic networks. *Bioinformatics*, 25(23):3158–3165.

Fischer, E. and Sauer, U. (2003). Metabolic flux profiling of *Escherichia coli* mutants in central carbon metabolism using GC-MS. *FEBS*, 270:880–891.

Fong, S. S., Nanchen, A., Palsson, B. O., and Sauer, U. (2006). Latent Pathway Activation and Increased Pathway Capacity Enable *Escherichia coli* Adaptation to Loss of Key Metabolic Enzymes. *Journal of Biological Chemistry*, 281(12):8024–8033.

Glaser, P. (2004). Mathematical models in microbial systems biology. *Current Opinion in Microbiology*, 7:513–518.

Gonzalez, J. E., Long, C. P., and Antoniewicz, M. R. (2016). Comprehensive analysis of glucose and xylose metabolism in *Escherichia coli* under aerobic and anaerobic conditions by 13C metabolic flux analysis. *Metabolic Engineering*.

Goodall, E. C. A., Robinson, A., Johnston, I. G., Jabbari, S., Turner, K. A., Lund, P. A., Cole, J. A., and Henderson, R. (2018). The Essential Genome of *Escherichia coli* K-12. *mBio*, 9(1):1–18.

Gudmundsson, S. and Thiele, I. (2010). Computationally efficient flux variability analysis. *BMC Bioinformatics*, 11(1):489.

Hanemaaijer, M., Röling, W. F. M., Olivier, B. G., Khandelwal, R. A., Teusink, B., and Bruggeman, F. J. (2015). Systems modeling approaches for microbial community studies: from metagenomics to inference of the community structure. *Frontiers in Microbiology*, 6:213.

Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular cell biology. *Nature*, 402:47–52.

Hilgetag, S. S. C. and Fell, J. H. W. D. A. (2002). Reaction routes in biochemical reaction systems : Algebraic properties , validated calculation procedure and example from nucleotide metabolism. *Mathematical Biology*, 45:153–181.

Hua, Q., Yang, C., Baba, T., Mori, H., and Shimizu, K. (2003). Responses of the Central Metabolism in *Escherichia coli* to Phosphoglucose Isomerase and Glucose-6-Phosphate Dehydrogenase Knockouts. *Journal of Bacteriology*, 185(24):7053–7067.

Husson, F., Josse, J., Le, S., and Mazet, J. (2016). *FactoMineR: Multivariate Exploratory Data Analysis and Data Mining.*

Kamp, A. V. and Klamt, S. (2014). Enumeration of Smallest Intervention Strategies in Genome-Scale Metabolic Networks. *PLoS Computational Biology*, 10(01).

Kiefer, P., Letisse, F., Kro, J., Soucaille, P., Wittmann, C., and Lindley, N. D. P. (2007). Response of the central metabolism of *Escherichia coli* to modified expression of the gene encoding the glucose-6-phosphate dehydrogenase. *FEBS Letters*, 581:3771–3776.

Kirschner, M. W. (2005). The Meaning of Systems Biology. *Cell*, 121:503–504.

Kitano, H. (2002). Computational systems biology. *Nature*, 420:206–210.

Klamt, S. (2006). Generalized concept of minimal cut sets in biochemical networks. *BioSystems*, 83:233–247.

Klamt, S. and Gilles, E. D. (2004). Minimal cut sets in biochemical reaction networks. *Bioinformatics*, 20(2):226–234.

Klamt, S., Regensburger, G., Gerstl, M. P., Jungreuthmayer, C., Schuster, S., Mahadevan, R., and Mu, S. (2017). From elementary flux modes to elementary flux vectors : Metabolic pathway analysis with arbitrary linear flux constraints. *PLoS Computational Biology*, 13(4):1–22.

Klamt, S. and Stelling, J. (2003). Two approaches for metabolic pathway analysis ?. *TRENDS in Biotechnology*, 21(2):64–69.

Klipp, J. S. W. L. E. (2009). Nested uncertainties in biochemical models. *IET Systems Biology*, 3(1):1–9.

Kondrashov, F. A., Koonin, E. V., Morgunov, I. G., Finogenova, T. V., and Kondrashova, M. N. (2006). Evolution of glyoxylate cycle enzymes in Metazoa : evidence of multiple horizontal transfer events and pseudogene formation. *Biology Direct*, 1:31.

Lewis, N. E., Hixson, K. K., Conrad, T. M., Lerman, J. A., Charusanti, P., Polpitiya, A. D., Adkins, J. N., Smith, R. D., Schramm, G., Purvine, S. O., Lopez-ferrer, D., Weitz, K. K., Eils, R., Ko, R., and Palsson, B. Ø. (2010). Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. *Molecular Systems Biology*, 6:390.

Liu, M., Feng, X., Ding, Y., Zhao, G., Liu, H., and Xian, M. (2015). Metabolic engineering of *Escherichia coli* to improve recombinant protein production. *Applied Microbiology Biotechnology*.

Llaneras, F. (2010). Which Metabolic Pathways Generate and Characterize the Flux Space ? A Comparison among Elementary Modes , Extreme Pathways and Minimal Generators. *Journal of Biomedicine and Biotechnology*, 2010.

Long, C. P. and Antoniewicz, M. R. (2014). Metabolic flux analysis of *Escherichia coli* knockouts : lessons from the Keio collection and future outlook. *Current Opinion in Biotechnology*, 28:127–133.

Macqueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *Math. Statist. and Prob.*, 1:281–297.

Maia, P., Rocha, M., and Rocha, I. (2016). *In Silico* Constraint-Based Strain Optimization Methods : the Quest for Optimal Cell Factories. *Microbiology and Molecular Biology Reviews*, 80(1):45–67.

Majewski, R. A. (1990). Simple Constrained-Optimization View of Acetate Overflow in *E. coli*. *Biotechnology and Bioengineering*, 35:732–738.

Mccloskey, D., Palsson, B. Ø., and Feist, A. M. (2013). Basic and applied uses of genome-scale metabolic network reconstructions of *Escherichia coli*. *Molecular Systems Biology*, 9(661):1–15.

Moffatt, B. A. and Ashihara, H. (2002). Purine and Pyrimidine Nucleotide Synthesis and Metabolism. In *The Arabidopsis Book / American Society of Plant Biologists*, number 1.

Mu, A. C. and Bockmayr, A. (2013). Systems biology Fast thermodynamically constrained flux variability analysis. *Bioinformatics*, pages 1–7.

Murarka, A., Clomburg, J. M., Gonzalez, R., and Gonzalez, R. (2010). Metabolic flux analysis of wild-type *Escherichia coli* and mutants deficient in pyruvate-dissimilating enzymes during the fermentative metabolism of glucuronate. *Microbiology*, 156:1860–1872.

Ng, C. Y., Preciat, G., Žagare, A., Chan, S. H. J., Aurich, M. K., Assal, D. C. E., Valcarcel, L. V., Apaolaza, I., and Ghaderi, S. (2018). Creation and analysis of biochemical constraint-based models: the COBRA Toolbox v3.0. *Nature Protocols*, 2(3):727–738.

Oliver, H. and Klamt, S. (2011). Computing complex metabolic intervention strategies using constrained minimal cut sets. *Metabolic Engineering*, 13:204–213.

Orth, J. D., Conrad, T. M., Na, J., Lerman, J. A., Nam, H., Feist, A. M., and Palsson, B. Ø. (2011). A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism. *Molecular Systems Biology*, 7(535):1–9.

Orth, J. D., Fleming, R. M. T., and Palsson, B. Ø. (2010a). Reconstruction and Use of Microbial Metabolic Networks: the Core *Escherichia coli* Metabolic Model as an Educational Guide. *EcoSalPlus*.

Orth, J. D., Thiele, I., and Palsson, B. Ø. (2010b). What is flux balance analysis ? *Nature Biotechnology*, 28(3):245–248.

Palsson, B. (2000). The challenges of *in silico* biology. *Nature*, 18:1147–1150.

Palsson, B. (2006). *Systems Biology: Properties of Reconstructed Networks*. Cambridge University Press, 1st edition.

Pandey, R., Kumar, N., Monteiro, G. A., Dasu, V., and Prazeres, D. M. F. (2018). Re-engineering of an *Escherichia coli* K-12 strain for the efficient production of recombinant human Interferon Gamma. *Enzyme and Microbial Technology*, 117(May):23–31.

Papin, J. A., Price, N. D., and Palsson, B. Ø. (2002). Extreme Pathway Lengths and Reaction Participation in Genome-Scale Metabolic Networks. *Genome Research*, 12:1889–1900.

Papin, J. A., Stelling, J., Price, N. D., Klamt, S., Schuster, S., and Palsson, B. O. (2004). Comparison of network-based pathway analysis methods. *TRENDS in Biotechnology*, 22(8):400–405.

Pfau, T., Christian, N., and Ebenho, O. (2011). Systems approaches to modelling pathways and networks. *Briefings in Functional Genomics*, 10(5):266–279.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*.

Raman, K. and Chandra, N. (2009). Flux balance analysis of biological systems : applications and challenges. *Briefings in Bioinformatics*, 10(4):435–449.

Reed, J. L., Vo, T. D., Schilling, C. H., and Palsson, B. O. (2003). An expanded genome-scale model of *Escherichia coli* K-12 (i JR904 GSM / GPR). *Genome Biology*, 4(9):1–12.

Ringnér, M. (2008). What is principal component analysis ?. *Nature Biotechnology*, 26(3):303–304.

Rokach, L. and Maimon, O. (2005). Clustering Methods. In *Data Mining and Knowledge Discovery Handbook*, pages 321–352.

Ruckerbauer, D. E., Hanscho, M., and Jungreuthmayer, C. (2013). Elementary flux modes in a nutshell : Properties , calculation and applications. *Biotechnology Journal*, 8:1–8.

Sariyar, B. and Ozde, F. (2005). Metabolic flux analysis of recombinant protein overproduction in Escherichia coli. *Biochemical Engineering Journal*, 22:167–195.

Schilling, C. H., Letsher, D., and Palsson, B. Ø. (2000). Theory for the Systemic Definition of Metabolic Pathways and their use in Interpreting Metabolic Function from a Pathway-Oriented Perspective. *Journal Theoritical Biology*, 203:229–248.

Schilling, C. H., Schuster, S., and Palsson, B. O. (1999). Metabolic Pathway Analysis : Basic Concepts and Scientific Applications in the Post-genomic Era. *Biotechnol. Prog.*, 15:296–303.

Schuster, S. and Hlgetag, C. (1994). On Elementary Flux Modes in Biochemical Reaction Systems At Steady State. *Journal of Biological Systems*, 2(2):165–182.

Smith, L. I. (2002). *A tutorial on Principal Components Analysis Introduction.*

Soh, K. C., Miskovic, L., and Hatzimanikatis, V. (2012). From network models to network responses: integration of thermodynamic and kinetic properties of yeast genome-scale metabolic netowrks. *FEMS Yeast Research*, 12:129–143.

Springi, T. G. and Wold, F. (1971). The Purification and Characterization of *Escherichia coli* Enolase. *Journal of Biological Chemistry*, 246(22):6797–6802.

Systems Biology Institute (2018). What is Systems Biology. `https://systemsbiology.org/about/what-is-systems-biology/`.

Szallasi, Z., Stelling, J., and Periwal, V. (2010). *System Modeling in Cellular Biology: From Concepts to Nuts and Bolts. The MIT Press.*

T. Trihn, C., Wlaschin, A., and Srienc, F. (2010). Elementary Mode Analysis: A Useful Metabolic Pathway Analysis Tool for Characterizing Celullar Metabolism. *Applied Microbiology*, 81(5):813–826.

Toya, Y., Nobuyoshi, I., Nakahigashi, K., Hirasawa, T., Tomita, M., Soga, T., and Shimizu, K. (2010). C-Metabolic Flux Analysis for Batch Culture of *Escherichia coli* and Its pyk and pgi Gene Knockout Mutants Based on Mass Isotopomer Distribution of Intracellular Metabolites. *Biotechnol. Prog.*, 26(4):975–992.

Usui, Y., Hirasawa, T., Furusawa, C., Shirai, T., and Yamamoto, N. (2012). Investigating the effects of perturbations to pgi and eno gene expression on central carbon metabolism in *Escherichia coli* using C metabolic flux analysis. *Microbial Cell Factories*, 11:87.

Vieira, J. (2015). *Development of pathway analysis based algorithms for strain optimization.* PhD thesis, Universidade do Minho, Portugal.

Waegeman, H., Lausnay, S. D., Beauprez, J., and Maertens, J. (2013). Increasing recombinant protein production in Escherichia coli K12 through metabolic engineering. *New Biotechnology*, 30(2):255–261.

Westerhoff, H. V. and Palsson, B. O. (2004). The evolution of molecular biology into systems biology. *Nature Biotechnology*, 22(10):1249–1252.

Wickham, H. (2009). *ggplot2 - Elegant graphics for data analysis.*

Widlak, W. (2013). High-Throughput Technologies in Molecular Biology. In *Molecular Biology*, pages 139–153.

Wolf, D. M. and Arkin, A. P. (2003). Motifs , modules and games in bacteria. *Current Opinion in Microbiology*, 6:125–134.

Yang, C., Hua, Q., Baba, T., Mori, H., and Shimizu, K. (2003). Analysis of *Escherichia coli* Anaplerotic Metabolism and Its Regulation Mechanisms From the Metabolic Responses to Altered Dilution Rates and Phosphoenolpyruvate Carboxykinase Knockout. *Biotechnology and Bioengineering*, 84(2):129–144.

Yang, S.-t., Liu, X., and Zhang, Y. (2007). Chapter 4 . Metabolic Engineering: Applications, Methods, and Challenges. In *Bioprocessing for Value-Added Products and Renewable Resources*, pages 73–118.

Yilmaz, L. S. and Walhout, A. J. M. (2017). Metabolic network modeling with model organisms. *Current Opinion in Chemical Biology*, 36:32–39.

Zhang, C. and Hua, Q. (2016). Applications of Genome-Scale Metabolic Models in Biotechnology and Systems Medicine. *Frontiers in Physiology*, 6:413.

Zhao, J., Baba, T., Mori, H., and Shimizu, K. (2004). Effect of zwf gene knockout on the metabolism of *Escherichia coli* grown on glucose or acetate. *Metabolic Engineering*, 6:164–174.

Zhu, J. and Shimizu, K. (2005). Effect of a single-gene knockout on the metabolic regulation in *Escherichia coli* for D -lactate production under microaerobic condition. *Metabolic Engineering*, 7:104–115.

Appendix A

Central Metabolism Model Reaction and Metabolite Lists

Table A.1: List of reactions and respective abbreviations used in the central metabolism model network. Adapted from supplementary data in Pandey *et al.* (2018).

Abbreviation	**Reaction Group**	**Reaction**
PTS	Glycolysis and Glucogenesis	Glc+PEP $\longrightarrow$ G6P+Pyr
PGM		G6P $\longrightarrow$ Glc+Pi
PGI		G6P $\longleftrightarrow$ F6P
PFK		F6P+ATP $\longrightarrow$ 2 T3P+ADP
FBA		2 T3P $\longrightarrow$ F6P + Pi
G3PD		T3P+ADP+Pi $\longleftrightarrow$ PG3+ATP+NADH
ENO		PG3$\longleftrightarrow$PEP
PYK		PEP+ADP$\longrightarrow$Pyr+ATP
PYC		PEP+CO2$\longrightarrow$OA
PEPCK		OA+ATP$\longrightarrow$PEP+CO2 +ADP+Pi
PDH		Pyr$\longrightarrow$AcCoA+CO2 +NADH
G6P1D	Pentose Phosphate Pathway	G6P$\longrightarrow$Gluc6P+NADPH
G1D		Glc$\longrightarrow$Gluc+NADH
GLUCK		Gluc+ATP$\longrightarrow$Gluc6P+ADP
6PGDH		Gluc6P$\longrightarrow$R5P+CO2+NADPH
RP3E		R5P$\longleftrightarrow$Xyl5P
R5PI		R5P$\longleftrightarrow$Rib5P
TKT1		Xyl5P+Rib5P$\longleftrightarrow$S7P+T3P
TALA1		Xyl5P+E4P$\longleftrightarrow$F6P+T3P
TALA2		T3P+S7P$\longleftrightarrow$F6P+E4P
ADH	Overflow Metabolism	AcCoA+NADH$\longleftrightarrow$Eth
ACK		AcCoA+ADP+Pi$\longleftrightarrow$ Ac+ATP

Table A.1 continued from previous page

Abbreviation	Reaction group	Reaction
PGDH	Entner Doudoroff Pathway	Gluc6P⟶Pyr+T3P
ICL	Glyoxylate Cycle	AcCoA+ ICit⟶Mal+Suc
MAL1	Malic Enzymes	Mal⟶Pyr+CO2 +NADH
MAL2		Mal⟶Pyr+CO2 +NADPH
CS	TCA Cycle	AcCoA+OA⟶Cit
ACONT		Cit⟶ICit
ICDH		ICit⟶αKG+CO2+NADH
AKGD		αKG⟶SucCoA+CO2+NADH
SUCOAS		SucCoA+Pi+ADP⟷Suc+ATP
SDH		Suc⟶Fum+FADH2
FUM		Fum⟶Mal
MDH		Mal⟶OA+NADH
PSP	Serine Family Amino Acids	PG3+Glu⟶Ser+αKG+NADH+Pi
GHMT		Ser+THF⟶Gly+MetTHF
STAC		Ser+AcCoA+H2S⟶Cys+Ac
ALATA	Alanine Family Amino Acids	Pyr+Glu⟶Ala+αKG
KAR		2Pyr+NADPH⟶Kval
VALTA		Kval+Glu⟶Val+αKG
LEUDH		Kval+AcCoA+Glu⟶Leu+αKG+NADH+CO2
RPPK	Histidine Family Amino Acids	R5P+ATP⟶PRPP+AMP
HISDH		PRPP+ATP+Gln⟶His+PRAIC+αKG+2Ppi+2NADH+Pi
ASPOX	Aspartic Acid Family Amino Acids	OA+Glu⟶Asp+αKG
ASPAS		Asp+Gln+ATP⟶Asn+Glu+AMP+Ppi
ASPK		Asp+ATP+NADPH⟶AspSa+ADP+Pi
DHDPS		AspSa+Pyr⟶DC
DHDPR		DC+NADPH⟶Tet
THPS		Tet+AcCoA+Glu⟶Ac+αKG+mDAP
DAPDC		mDAP⟶Lys+CO2
HOMD		AspSa+NADPH⟶HSer
HOMSK		Hser+ATP⟶Thr+ADP+Pi
THRDH		α Thr+Pyr+NADPH+Glu⟶Ile+αKG+NH3 +CO2
HOMST		AcCoA+Cys+HSer+H2S+MTHF⟶Met+Pyr+2Ac+NH3+THF

Table A.1 continued from previous page

Abbreviation	Reaction group	Reaction
CHORS	Aromatic Family	2PEP+E4P+ATP+NADPH⟶Chor+ADP+4Pi
CHORM		Chor+Glu⟶Phe+αKG+CO2
PRPDH		Chor+Glu⟶Tyr+αKG+CO2 +NADH
GLUTS	Glutamic Acid Family Amino Acids	αKG+NH3+NADPH⟶Glu
GLUTST		Glu+ATP+NH3 +⟶Gln+ADP+Pi
PYRRDH		Glu+ATP+2NADPH⟶Pro+ADP+Pi
ORNTA		2Glu+AcCoA+ATP+NADPH⟶Orn+αKG+Ac+ADP+Pi
ORNCT		Orn+CaP⟶Citr+Pi
ARGSS		Citr+Asp+ATP⟶Arg+Fum+AMP+PPi
APPRT	Nucleotide Synthesis	PRPP+2Gln+Asp+CO2+Gly+4ATP+F10THF⟶ 2Glu+PPi+4ADP+4Pi+THF+PRAIC+Fum
PRISC		PRAIC+F10THF⟶IMP+THF
I5MPDH		IMP+Gln+ATP⟶NADH+GMP+Glu+AMP+PPi
GUAK		GMP+ATP⟶GDP+ADP
GDPK		ATP+GDP⟷ADP+GTP
DATPK		ATP+NADPH⟶dATP
DGTPK		GDP+ATP+NADPH⟶ADP+dGTP
ADSUCS		IMP+GTP+Asp⟶GDP+Pi+Fum+AMP
ADK		AMP+ATP⟶2ADP
ASPCMT		PRPP+Asp+CaP⟶UMP+NADH+PPi+Pi+CO2
UMPK		UMP+ATP⟶ADP+UDP
UDPK		UDP+ATP⟶ADP+UTP
CTPS		UTP+NH3+ATP⟶CTP+ADP+Pi
DCTPK		ATP+NADPH+CDP⟶dCTP+ADP
CDPK		CDP+ATP⟷CTP+ADP
THYMK		UDP+MetTHF+2ATP+NADPH⟶dTTP+DHF+2ADP+PPi
DHFR	One Carbon Units	DHF+NADPH⟶THF
MTHFT		MetTHF+CO2+NH3+NADH⟶Gly+THF
MTHFR		MetTHF+NADPH⟶MTHF
MTHFD		MetTHF⟶MeTHF+NADPH
MTHFC		MeTHF⟶F10THF
TRANSH1	Transhydrogenase Reactions	0.25ATP+NADH⟶NADPH+0.25ADP+0.25Pi
TRANSH2		NADPH⟶NADH

Table A.1 continued from previous page

Abbreviation	Reaction group	Reaction
ATPS1	Electron Transport	NADH+0.5O2++2ADP+2Pi⟶2ATP
ATPS2		FADH2+ADP+Pi+0.5O2⟶ATP
GL3PD	Fatty Acid Synthesis	T3P+NADPH⟶GL3P
FAS1		7 AcCoA + 6 ATP + 12 NADPH ⟶ C14:0 + 6 ADP + 6 Pi
FAS2		7 AcCoA + 6 ATP + 11 NADPH ⟶ C14:0 + 6 ADP + 6 Pi
FAS3		8.2 AcCoA + 7.2 ATP + 14 NADPH ⟶ FA+ 7.2 ADP + 7.2 Pi
FAS4		2 ATP + CO2 + Gln ⟶ CaP + Glu + 2 ADP + Pi
GLUTT	Other Biomass Components	F6P + Gln + AcCoA + UTP ⟶ UDPNAG + Glu + PPi
GLCNACS		PEP + NADPH + UDPNAG ⟶ UDPNAM + Pi
CMPKDOS		RL5P + PEP + CTP ⟶ CMPKDO + PPi + 2 Pi
PPDSDC		Ser + CTP + ATP ⟶ CDPEtN + ADP + PPi+ CO2
PGM2		G6P ⟶ G1P
UTPG1PUT		UTP + G1P ⟶ UDPGlc + PPi
BiomassProduction	Biomass	0.594 Ala + 0.198 Arg + 0.143 Asn + 0.284 Asp + 0.060 Cys + 0.272 Gln + 0.367 Glu + 0.495 Gly + 0.086 His + 0.288 Ile + 0.368 Leu + 0.342 Lys + 0.118 Met + 0.059 Orn + 0.175 Pro + 0.304 Ser + 0.239 Thr + 0.335 Val +0.17 Phe + 0.13 Tyr + 0.05 Trp + 0.136 UTP + 0.126 CTP + 0.203 GTP + 0.0246 dATP + 0.0254 dGTP + 0.0254 dCTP + 0.0246 dTTP + 0.083 GL3P + 0.0238 C14:0 + 0.0238 C14:1 + 0.15 FA + 0.095 UDPNAG + 0.095 UDPNAM + 0.111 UDPGlc + 0.154 + G1P+ 0.0235 CMPKDO + 0.0235 CDPEtN + 22.738 ATP ⟶ 1g Biomass + 22.738 ADP + 22.738 Pi
ATPM	Maintenance	ATP⟶ADP+Pi
CO2_e	Transport Reactions	CO2⟷exp
NH3_e		Imp⟷NH3
H2S_e		2ATP+4NADPH⟶AMP+ADP+H2S+PPi+Pi
PPI		PPi⟶2Pi
Pi_e		Imp⟷Pi

Table A.1 continued from previous page

Abbreviation	Reaction group	Reaction
AA_e		Ser+PRPP+Gln+Chor⟶Trp+Glu+CO2+Pyr+T3P+Ppi
GLC_e		Imp⟶Glc
O2_e		Imp⟶O2
ETH_e		Eth⟶exp
AC_e		Ac⟶exp
Biomass_e	Biomass Synthesis	Biomass⟶exp

Table A.2: List of metabolites and respective abbreviations used in the central metabolism model network. Adapted from supplementary data in Pandey *et al.* (2018).

Abbreviation	Metabolite
Ac	Acetate
AcCoA	Acetyl coenzyme A
Actn	Acetoin
ADP	Adenosine 5' -diphosphate
Ala	L-Alanine
AMP	Adenosine 5'-monophosphate
Arg	L-Arginine
Asn	L-Asparagine
Asp	L-Aspartate
AspSa	Aspartate semialdehyde
ATP	Adenosine 5'-triphosphate
C14:0	Myristic acid
C14:1	Hydroxymyristic acid
CaP	Carbamoyl-phosphate
CDP	Cytidine 5'-diphosphate
CDPEtN	CDP-ethanolamine
Cit	Citrate
Citr	Citruline
Chor	Chorismate
CMP	Cytidine 5'-monophosphate
CMPKDO	CMP-3-deoxy-D-manno-octulosonic acid
CO2	Carbon dioxide
CTP	Cytidine 5'-triphosphate
Cys	L-Cysteine
dATP	2' -Deoxy-ATP
dCTP	2' -Deoxy-CTP

Table A.2 continued from previous page

Abbreviation	Metabolite
dGTP	2' -Deoxy-GTP
dTTP	2' -Deoxy-TTP
DC	L,2,3 dihydrodipicolinate
DHF	7,8-Dihydrofolate
E4P	Erythrose 4-phosphate
Eth	Ethanol
F10THF	N10 -Formyl-THF
F6P	Fructose 6-phosphate
FADH	Flavine adenine dinucleotide (reduced)
Fum	Fumarate
G1P	Glucose 1-phosphate
G6P	Glucose 6-phosphate
GDP	Guanosine 5'-diphosphate
GL3P	Glycerol 5'-phosphate
Glc	Glucose
Gln	L-Glutamine
Glu	L-Glutamate
Gluc	Gluconate
Gluc6P	Gluconate 6-phosphate
Glx	Glyoxylate
Gly	L-Glycine
GMP	Guanosine 5'-monophosphate
GTP	Guanosine 5'-triphosphate
H2S	Hydrogen sulfide
His	L-Histidine
HSer	Homoserine
ICit	Isocitrate
Ile	L-Isoleucine
IMP	Inosine monophosphate
aKG	a-ketoglutarate
Kval	Ketovaline
Leu	L-Leucine
Lys	L-Lysine
Mal	Malate
mDAP	meso-Diaminopimelate
Met	L-Methionine
MeTHF	N5-N10-methenyl-THF
MetTHF	N5-N10-methylene-THF
MTHF	N5-methyl-THF
NADH	Nicotinamide adenine dinucleotide (reduced)

Table A.2 continued from previous page

Abbreviation	Metabolite
NADPH	Nicotinamide adenine dinucleotide phosphate (reduced)
NH3	Ammonia
OA	Oxalacetate
Orn	Ornithine
PA	Fatty acids
PEP	Phosphoenolpyruvate
PG3	Glycerate 3-phosphate
Phe	L-Phenylalanine
Pi	Inorganic orthophosphate
PPi	Inorganic pyrophosphate
PRAIC	5'-Phosphoribosyl-4-carboxamide-5-aminoimidazole
Pro	L-Proline
PRPP	5-Phospho-D-ribosylpyrophosphate
Pyr	Pyruvate
R5P	Ribulose 5-phosphate
Rib5P	Ribose 5-phosphate
S7P	Sedoheptulose-7-phosphate
Ser	L-Serine
Suc	Succinate
SucCoA	Succinate coenzyme A
Xy15P	Xylulose 5-phosphate
Tet	L,2,3,4,5 Tetrahydrodipicolinate
T3P	Triose 3-phosphate
THF	Tetrahydrofolate
Thr	L-Threonine
Trp	L-Tryptophan
Tyr	L-Tyrosine
UDP	Uridine 5'-diphosphate
UDPGlc	UDP-glucose
UDPNAG	UDP-N-acetyl-glucosamine
UDPNAM	UDP-N-acetyl-muramic acid
UMP	Uridine 5'-monophosphate
UTP	Uridine 5'-triphosphate
Val	L-Valine

Appendix B

Biomolecules Composition

Table B.1: Nucleotide composition of pET28a(+) sequence (that already accounts for the resistance marker sequence) added to IFNγ nucleotidic sequence (NCBI database reference sequence number AB451324.1).

Nucleotide	**Code**	**MW (g/mol)**	**# in pET28a**	**# pET28a dS**	**% Nucleotidic**	**MW in pET28a (g/mol)**	**mmole/g pET28a**
dATP	A	331.2	1 446	2 892	24.63	957 830.4	0.7591
dTTP	T	322.2	1 395	2 790	23.76	898 938	0.7324
dGTP	G	347.2	1 551	3 102	26.42	1 077 014.4	0.8143
dCTP	C	307.2	1 478	2 956	25.18	9 08 083.2	0.7759
	Total		5 870	11 740	100	3809630	

Table B.2: Amino acid composition of human interferon gamma fused to an hexa-histidine affinity tag (NCBI database reference sequence number NP_000610.2 - Interferon gamma precursor [homo sapiens]).

Amino acid (AA)	**Code**	**MW (g/mol)**	**MW - MW(H_2O)**	**# in IFNγ**	**% AA**	**MW in IFNγ (g/mol)**	**mmole/g IFNγ**
Alanine	A	89.09	71.08	10	5.29	710.80	0.4587
Arginine	R	174.19	156.18	9	4.76	1405.62	0.4128
Asparagine	N	132.11	114.10	10	5.29	1141.00	0.4587
Aspartic Acid	D	133.10	115.09	10	5.29	1150.90	0.4587
Cysteine	C	121.15	103.14	3	1.59	309.42	0.1376
Glutamic acid	E	147.13	129.12	9	4.76	1162.08	0.4128
Glutamine	Q	146.14	128.13	10	5.29	1281.30	0.4587
Glycine	G	75.06	57.05	10	5.29	570.50	0.4587
Histidine	H	155.15	137.14	9	4.76	1234.26	0.4128
Isoleucine	I	131.17	113.16	9	4.76	1018.44	0.4128
Leucine	L	131.17	113.16	15	7.94	1697.40	0.6880
Lysine	K	146.18	128.17	21	11.11	2691.57	0.9633
Methionine	M	149.20	131.19	7	3.70	918.33	0.3211
Phenylalanine	F	165.19	147.18	11	5.82	1618.98	0.5046
Proline	P	115.13	97.12	3	1.59	291.36	0.1376
Serine	S	105.09	87.08	19	10.05	1654.52	0.8715
Threonine	T	119.12	101.11	6	3.17	606.66	0.2752
Typtophan	W	204.22	186.21	1	0.53	186.21	0.0459
Tyrosine	Y	181.19	163.18	7	3.70	1142.26	0.3211
Valine	V	117.14	99.13	10	5.29	991.3	0.4587
		Total		189	100	21800.92	

Table B.3: Amino acid composition of plasmid resistance marker phosphotransferase (NCBI database reference sequence number WP_000018329 - aminoglycoside O-phosphotransferase APH(3')-Ia [Bacteria] (kanR)).

Amino acid (AA)	**Code**	**MW (g/mol)**	**MW - MW(H_2O)**	**# in kanR**	**% AA**	**MW in kanR (g/mol)**	**mmole/g kanR**
Alanine	A	89.09	71.08	15	5.54	1066.20	0.4842
Arginine	R	174.19	156.18	16	5.90	2498.88	0.5165
Asparagine	N	132.11	114.10	15	5.54	1711.50	0.4842
Aspartic Acid	D	133.10	115.09	25	9.23	2877.25	0.8070
Cysteine	C	121.15	103.14	5	1.85	515.70	0.1614
Glutamic acid	E	147.13	129.12	13	4.80	1678.56	0.4196
Glutamine	Q	146.14	128.13	10	3.69	1281.30	0.3228
Glycine	G	75.06	57.05	17	6.27	969.85	0.5488
Histidine	H	155.15	137.14	7	2.58	959.98	0.2260
Isoleucine	I	131.17	113.16	13	4.80	1471.08	0.4196
Leucine	L	131.17	113.16	29	10.70	3281.64	0.9361
Lysine	K	146.18	128.17	12	4.43	1538.04	0.3874
Methionine	M	149.20	131.19	8	2.95	1049.52	0.2582
Phenylalanine	F	165.19	147.18	16	5.90	2354.88	0.5165
Proline	P	115.13	97.12	15	5.54	1456.80	0.4842
Serine	S	105.09	87.08	16	5.90	1393.28	0.5165
Threonine	T	119.12	101.11	10	3.69	1011.10	0.3228
Typtophan	W	204.22	186.21	6	2.21	1117.26	0.1937
Tyrosine	Y	181.19	163.18	7	2.58	1142.26	0.2260
Valine	V	117.14	99.13	16	5.90	1586.08	0.5165
		Total		271	100	30979.17	

Appendix C

Hierarchical Clustering Analysis

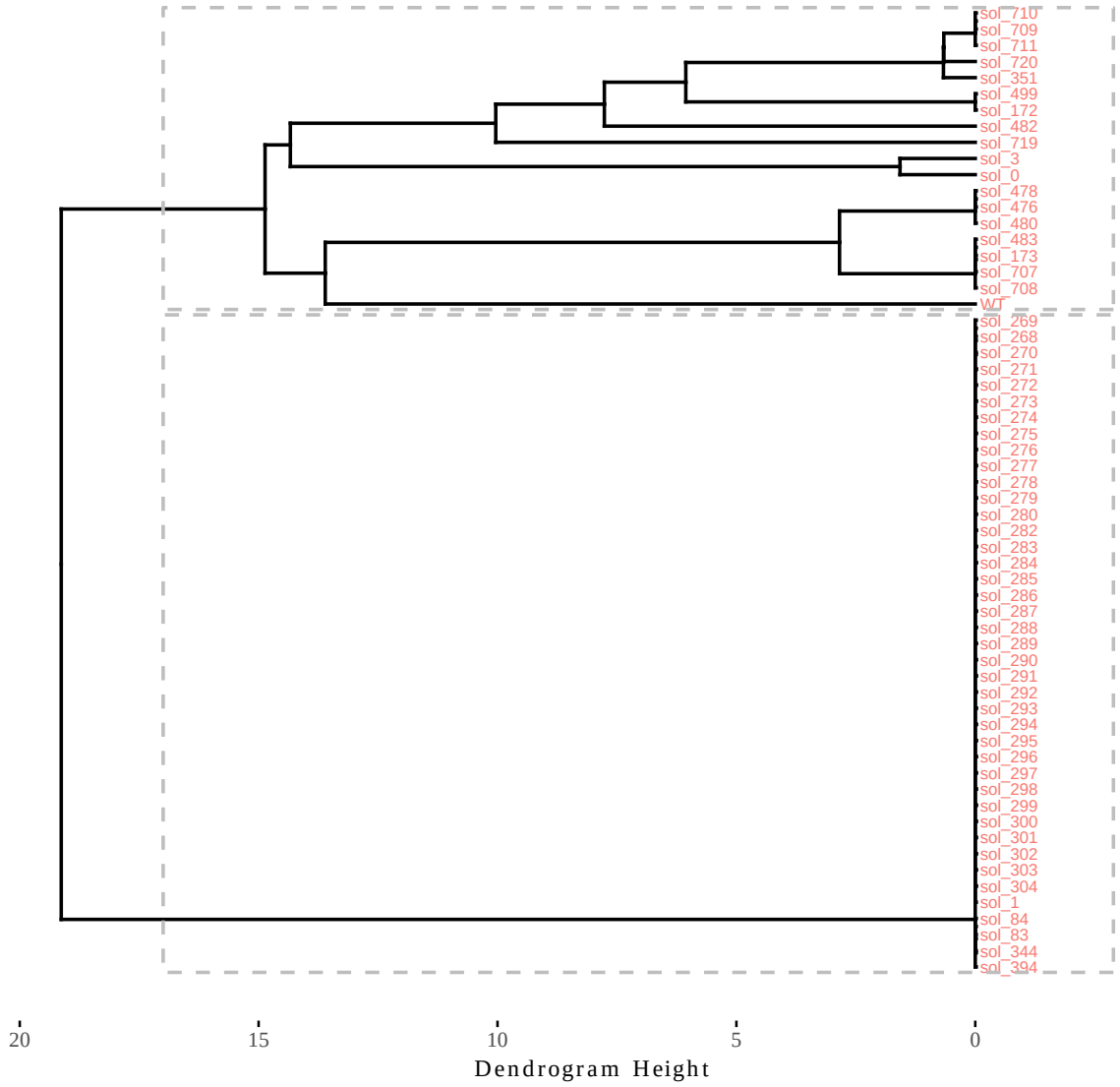

Figure C.1: Model CMM_A HCA full dendrogram obtained using single linage with Euclidean distance. The colour in solution labelling refers to the colours used for the respective enumeration problems.

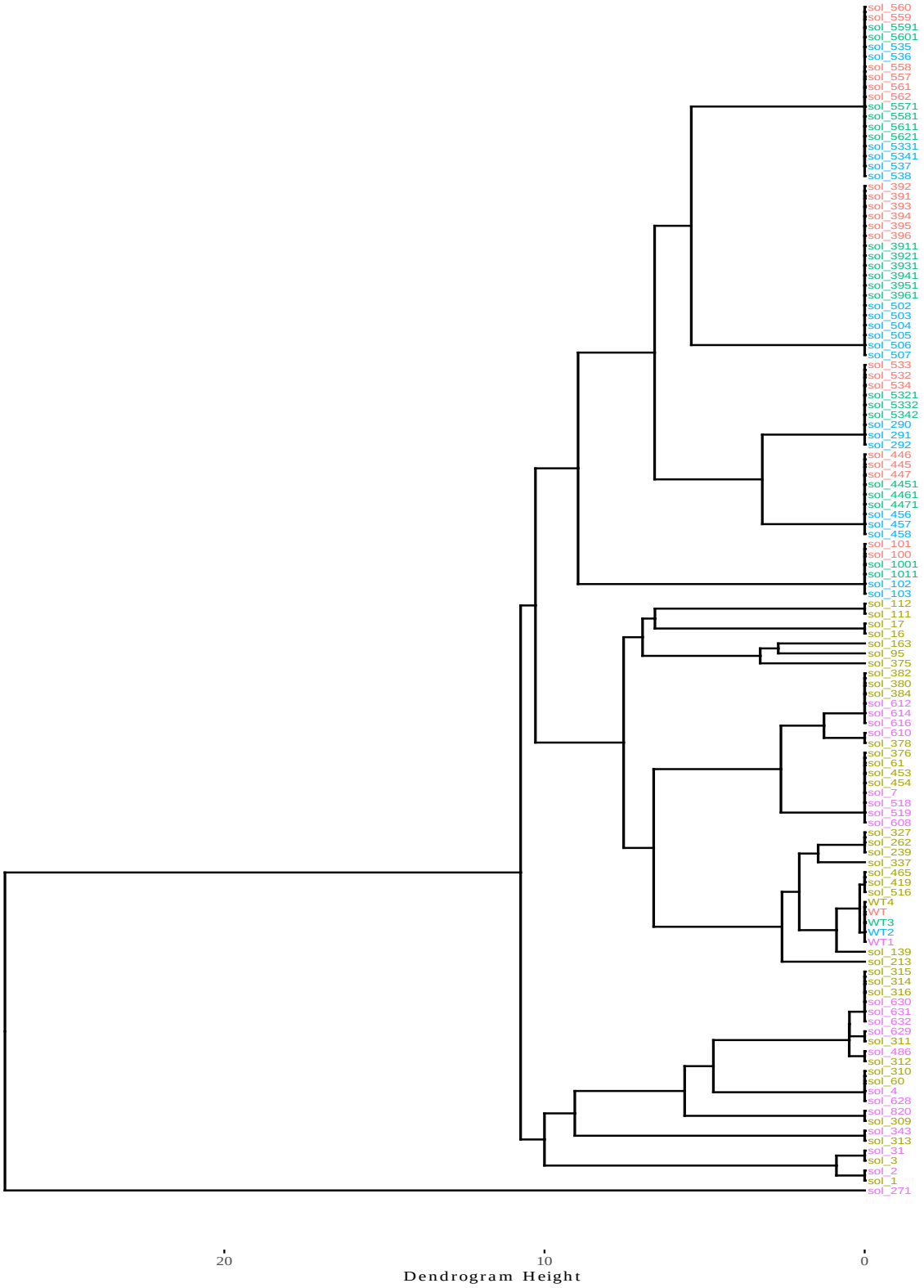

Figure C.2: Model CMM_B HCA full dendrogram obtained using single linage with Euclidean distance. The colour in solution labelling refers to the colours used for the respective enumeration problems.

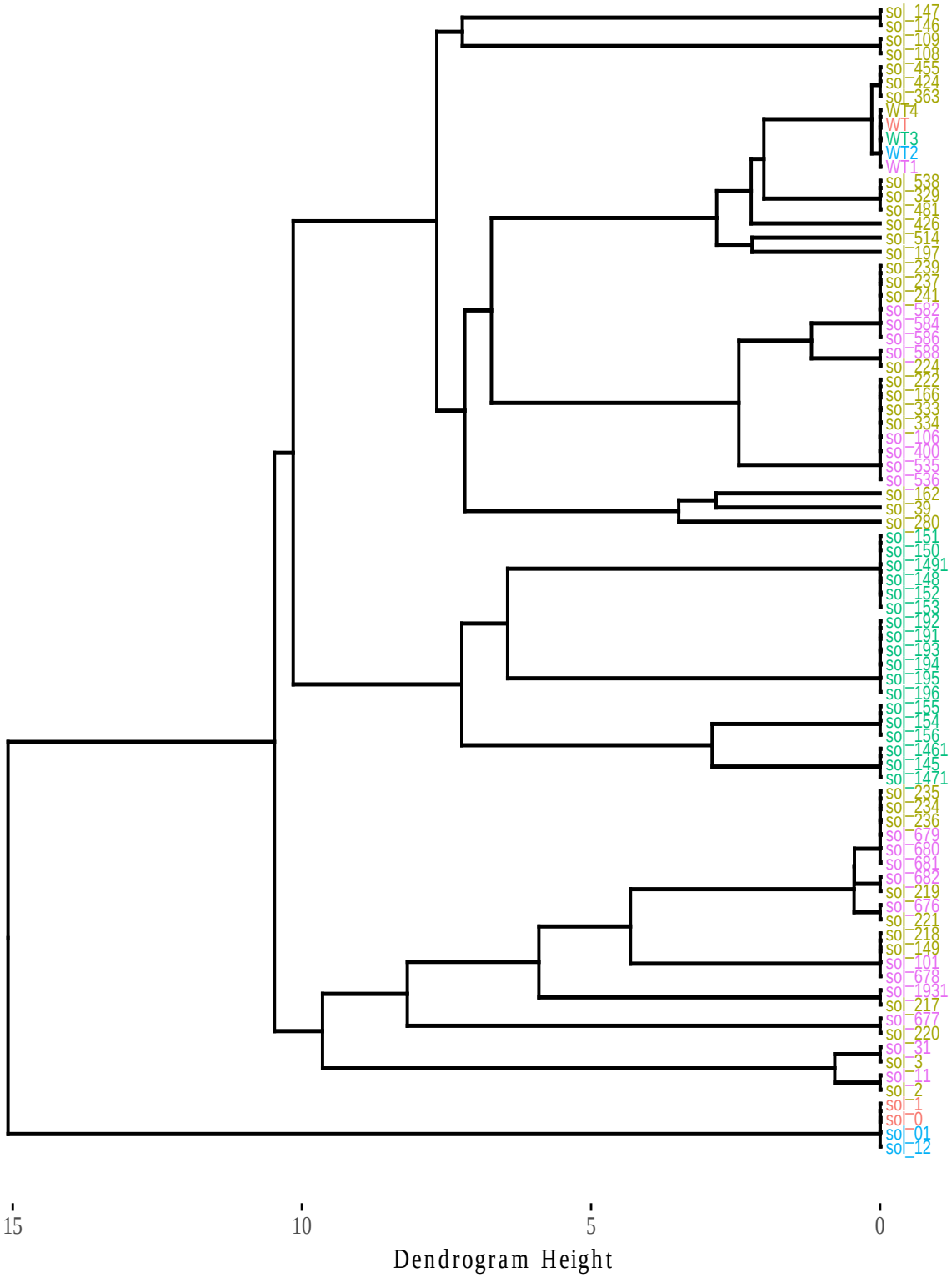

Figure C.3: Model CMM_C HCA full dendrogram obtained using single linage with Euclidean distance. The colour in solution labelling refers to the colours used for the respective enumeration problems.

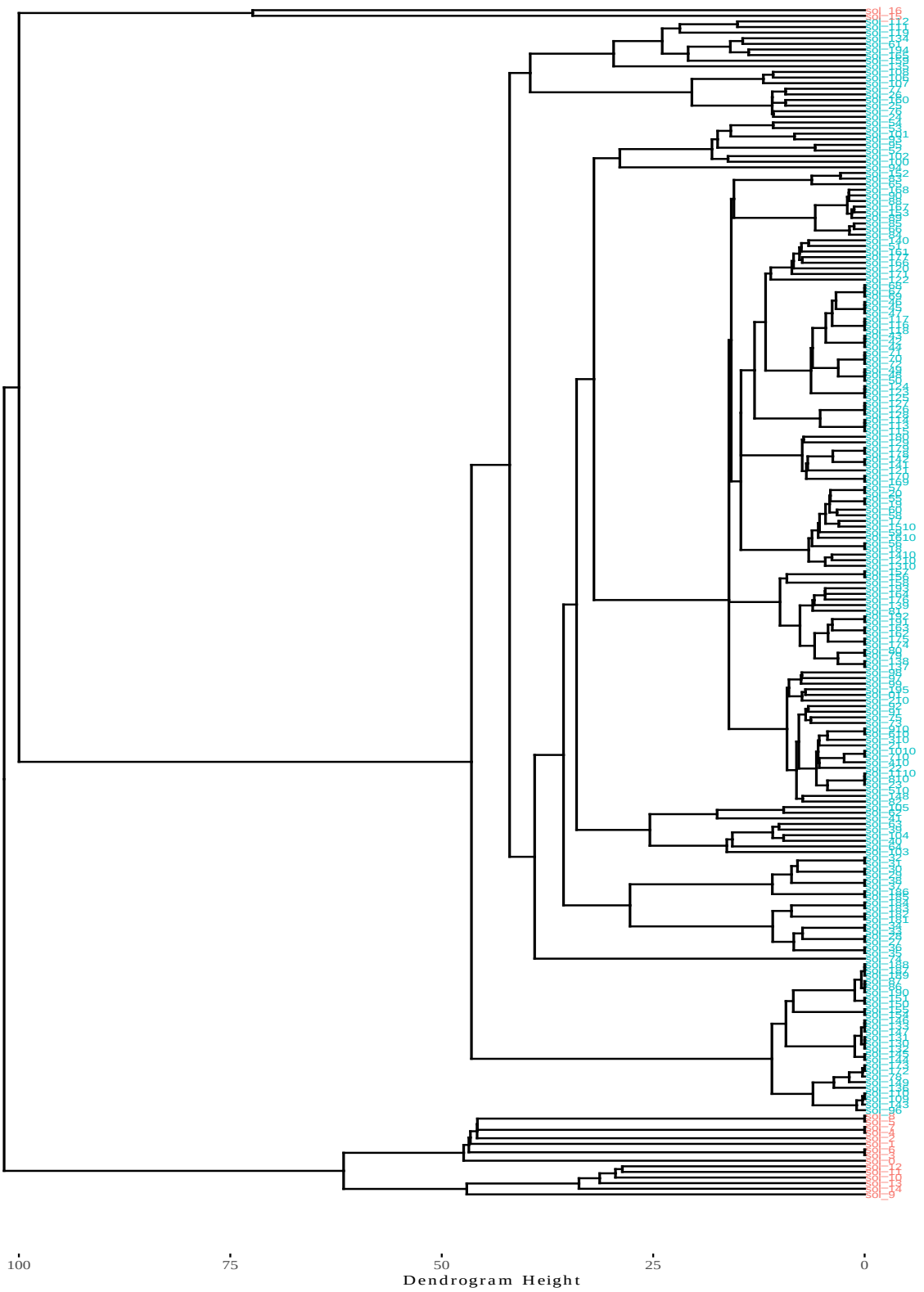

Figure C.4: Model GSM_B HCA full dendrogram obtained using single linage with Euclidean distance. The colour in solution labelling refers to the colours used for the respective enumeration problems.

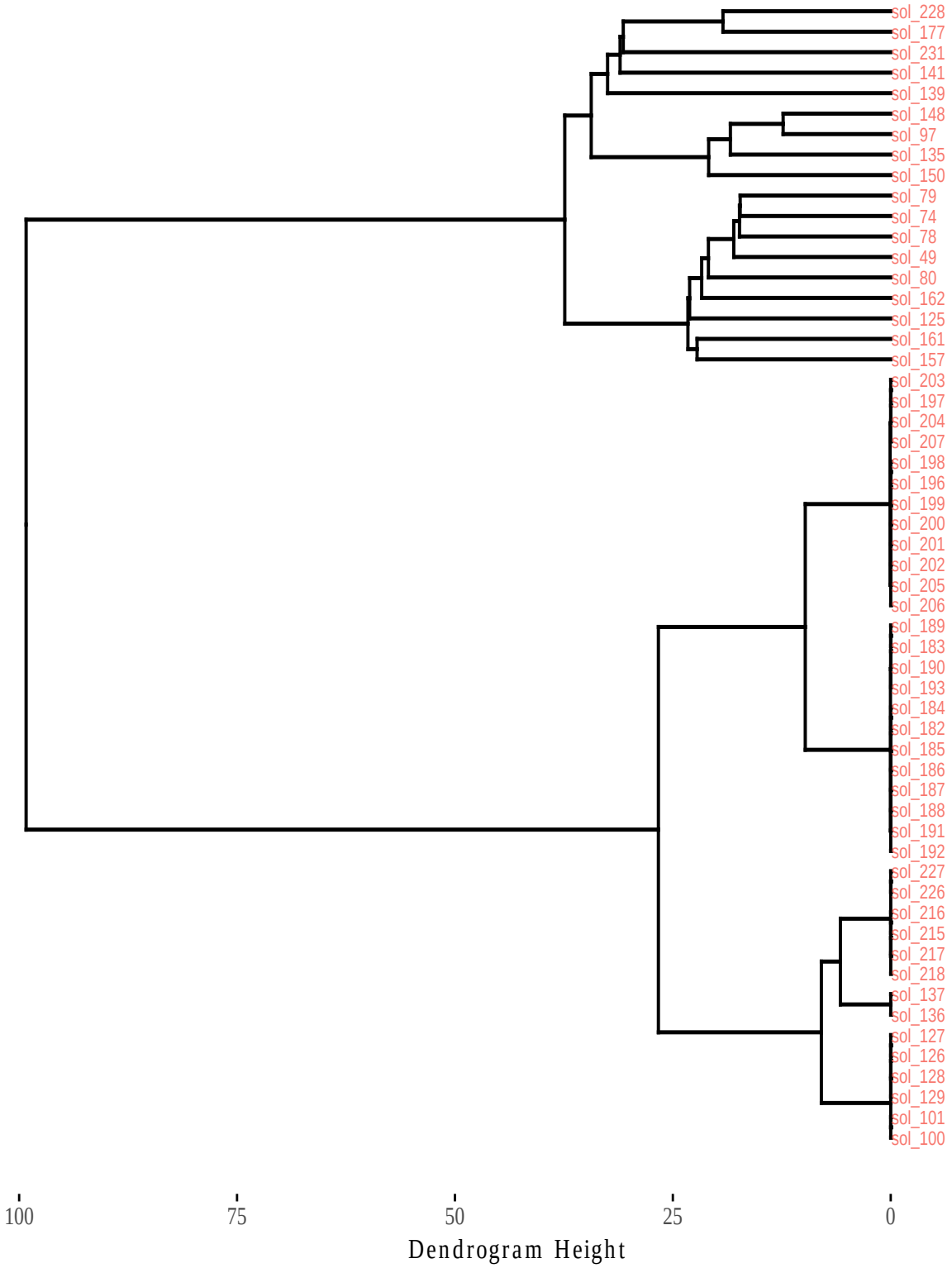

Figure C.5: Model GSM_C HCA full dendrogram obtained using single linage with Euclidean distance. The colour in solution labelling refers to the colours used for the respective enumeration problems.