NTNU – Trondheim
Norwegian University of
Science and Technology

Modelling and Analysis of Osmotic Stress in Escherichia Coli

Eivind Bøe Drejer

Biotechnology (5 year)
Submission date: May 2015
Supervisor: Eivind Almaas, IBT

Norwegian University of Science and Technology
Department of Biotechnology

Acknowledgements

Since starting the work on this thesis, I have had the help of many people close to me. Their support and guidance have been paramount to the completion of this project, and I am deeply grateful for all the help I have been given.

I would like to thank my research group, and in particular my supervisor, Professor Eivind Almaas, for your guidance and advice. I will always remember our discussions on the fine details of metabolic modelling, science and life. You have been a great mentor to me, and I have grown immensely under your supervision these past two years.

To Dr. József Baranyi, Dr. Aline Métris, Susie George and Daniel Marin at the Computational Microbiology Research Group at the Institute for Food Research in Norwich: Thank you for all the help you've given me on this project. I greatly enjoyed my stays in Norwich, and I will always look back on them fondly. A special thank you goes to Aline, who has been a source of good discussions and ideas.

To all my friends, both in Trondheim and back home, I would like to say thank you for all the memories and the good times we have had. You have made my time as a student the best years of my life, and I feel very lucky to have all of you.

When times have been tough, I have always been able to rely on my family for support. I could not ask for a better family, and I look forward to spending more time with you in the years to come. Thank you for always being there for me, and for taking such good care of me.

Finally, I would like to thank the love of my life, Eva. You are always there when I need you, and I could not have done this without you. Thank you for all the help you've given me in writing my thesis, boosting my morale when it was at its lowest and motivating me to do my best.

Sammendrag

Det har blitt påvist at mikroorganismers vekstrater påvirkes av flere abiotiske stresskilder, slik som osmotisk stress.[1,2,3] Forståelse for tilpasningene som skjer i mikroorganismer under osmotisk stress er viktig, da det kan bistå utviklingen av virkemidler og metoder for konservering av mat, som er en vanlig smittekilde for patogener hos mennesker.[4,5]

En av de vanligste patogene organismene som infiserer mennesker er *Escherichia coli* (*E. coli*).[6] I 2004 rapporterte "The Center for Disease Control and Prevention" i USA om to utbrudd av patogen *E. coli*, der smitte–kilden i begge tilfeller var mat.[7,8] I U-land er akutt diaré den nest vanligste dødsårsaken for spedbarn, der *E. coli* er en av de vanligste kildene til sykdommen.[9] For å motvirke smitte av *E. coli* i mennesker er det derfor viktig å utvikle metoder som kan modellere og predikere hvordan denne organismen reagerer på eksterne stressfaktorer.

Målet med denne masteroppgaven var å undersøke metabolismen til *E. coli* som ble utsatt for varierende grader av osmotisk stress. For å oppnå dette ble det opprettet et
samarbeid med "Institute for Food Research" (IFR) i Norwich, Storbritannia. Gjennom samarbeid med "Computational Microbiology Research Group" ved IFR ble det tatt målinger av genuttrykk i *E. coli* utsatt for varierende grader av osmotisk stress. Disse målingene ble deretter analysert ved bruk av metabolsk modellering.

Osmotisk stress er et komplekst fenomen, og det ble i løpet av prosjektet nødvendig å utvikle en ny metode for integrering av metabolske modeller og genuttrykks–målinger, kalt "Metabolic Flux Distribution by Translational Efficiency and Enzyme Kinetics" (MUTE). MUTE er i stand til å komme med prediksjoner om endringer i metabolismen til organismer basert på genuttrykks–målinger, enzymkinetikk og translasjonseffektivitet. Metoden ble vist å være mer sensitiv for forandringer i genuttrykk enn andre sammenlignbare metoder som "Metabolic Adjustment by Differential Expression" (MADE). Dette resulterte i nye prediksjoner for forandringer i metabolismen til *E. coli* utsatt for osmotisk stress. En interessant nyvinning hos MUTE er detaljnivået metoden opererer med, der proteinkonsentrasjons-prediksjoner ligger i samme størrelsesorden som det empiriske data rapporterer.[10]

Abstract

Microorganisms are known to be affected by stresses such as osmotic stress by reducing their growth rate.[1,2,3] Understanding the mechanisms behind this are important, as it can aid in the development of new methods of conserving food – a common source of pathogen infection for humans.[4,5]

One of the most common pathogenic infections for humans is *Escherichia coli* (*E. coli*).[6] In 2014, the Center for Disease Control and Prevention in the USA reported two outbreaks of pathogenic *E. coli*, both transmitted through food.[7,8] In developing countries, acute diarrhea is the second most common cause of infant death, and infection by *E. coli* is one of the most common sources.[9] In order to effectively combat *E. coli* infection in humans, it is important that accurate methods for predicting the organism's response to external stresses are developed.

The goal of this master thesis was to investigate the metabolism of *E. coli* under osmotic stress. In order to accomplish this, the project was set up as a collaboration with the Institute for Food Research (IFR) in Norwich, United Kingdom. Through collaboration with the Computational Microbiology Research Group at IFR, gene expression data for *E. coli* growing under different states of osmotic stress was collected and analyzed using metabolic modelling.

The complex nature of osmotic stress required the development of a new method, dubbed Metabolic Flux Distribution by Translational Efficiency and Enzyme Kinetics (MUTE). MUTE is able to predict changes in metabolic flux based on gene expression data, translation efficiencies and enzyme kinetics. MUTE was shown to increase the sensitivity to expression data compared to other methods such as “Metabolic Adjustment by Differential Expression” (MADE), resulting in new predictions on metabolic changes during osmotic stress in *E. coli.* Another novelty of MUTE is its level of detail, where enzyme concentration predictions are levels reported by empirical data.[10]

Contents

List of Abbreviations

E. coli	*Escherichia coli*
MUTE	Metabolic flUx distribution by Translational efficiency and Enzyme kinetics
MOMENT	MetabOlic Modeling with Enzyme kiNeTics
GLPK	GNU Linear Programming Kit
CPLEX	IBM ILOG CPLEX Optimization Studio
GRN	Gene Reaction Network
FBA	Flux Balance Analysis
GX-FBA	Gene–eXpression Flux Balance Analysis
BRENDA	Braunschweig Enzyme Database
SABIO–RK	System for the Analysis of Biochemical Pathways – Reaction Kinetics
ORF	Open Reading Frame
*i*AF1260	Metabolic reconstruction of *E. coli* accounting for 1260 ORFs
*i*JO1366	Metabolic reconstruction of *E. coli* accounting for 1366 ORFs
SBML	Systems Biology Markup Language
COBRA	COnstraint Based Reconstruction and Analysis
OMICS	Grouping term for large sets of biological data
TIGER	Toolbox for Integrating Genome-scale metabolism, Expression and Regulation
ELF	Executable and Linkable Format
	BiGG Biochemical Genetic and Genomic knowledgebase

Chapter 1

Introduction

A long standing strategy in the quest to understand life at a cellular and molecular level has been a reductionist one, separating complex processes and systems into simpler ones to allow for their analysis.[11] The fruits of this labour can not be understated, as it has provided us with a wealth of new information that now allows us to accomplish feats in life sciences that we were only able to dream of no more than 20 years ago, such as the sequencing of the human genome.[12]

The exponential growth in information generation made possible by "next generation" biology tools has highlighted the need for a complimentary approach to reductionism.[13,14,15] As the details governing life are elucidated, it has become evident that many systems can only truly be understood by looking at their emergent properties when viewing the simplified systems as a whole.[16,17] One of the ways in which this can be accomplished is by using tools from the field of network analysis, which is routinely applied in various other fields such as computer science, sociology and physics.[18] This approach allows both visual and computer aided identification of novel clusters of connected components, making possible the generation of new hypotheses and their subsequent experimental testing.[11]

Systems biology is in part built on the progress made in network theory and routinely applies it to visualize and analyze biological systems.[11,16] Analysis of gene expression has allowed for construction of gene regulatory networks (GRNs), which help shed light on the processes governing a cells' behaviour under a wide array of conditions. The access to GRNs for several organisms has helped develop a sub–field of systems biology; genetic circuits, where potentially novel combinations of genes and regulatory elements are constructed in order to create novel biological functions.[19] Metabolic modelling has also benefited from the network approach to understanding life.

The data and insight gained by next–generation tools and bioinformatics has, for several organisms, resulted in compiled lists of most biochemical reactions and pathways for several organisms, such as humans.[20] Knowing the reactions that happen inside a cell is useful in itself, and pairing it with modern optimization theory and constraint based analysis enables modelling of the metabolism entire cells.[21,22,23,24] This type of modelling predicts the "flow" of matter, or flux, through each reaction or pathway of the cell in question. The list of organisms whose metabolism has been modelled is rapidly growing, and several different approaches are used to construct the models, differing both in the data used to impose restrictions on the system and the methods for optimizing and solving for the fluxes of the system.[24,21]

By using information contained in the GRNs to impose retrictions on system behaviour and flux, it is possible to remove "extreme" pathways which are only active in a minor set of circumstances, resulting in a flux distribution that is more realistic for "normal" cell states.[22] This is known as shrinking the solution space of the model.[24] A promising next step for constraint-based metabolic models is to integrate gene expression data to shrink the solution space ever further and improve upon the accuracy of predictions.[24]

The availability of gene expression data has led to their inclusion in constraint-based metabolic models. Some models use gene expression data as a "boolean switch", where they are used to determine whether a given pathway or reaction is on or off.[25] This inclusion of expression data allows for a more accurate picture of cell metabolism to be modelled, however, the boolean nature of the expression data's incorporation may be too simplistic for accurate modelling. One method adressing this was Gene Expression Flux Balance Analysis (GX-FBA), in which gene expression levels were used to set lower and upper bounds on the flux of each pathway, resulting in more accurate predictions[23].

Systems biology methods are maturing at a rapid pace, and the complexity and level of details of models and predictions is constantly increasing. Making predictions about cell metabolism and other facets of cells is now common practice in many research groups, and the applications of metabolic modelling are spreading to most research areas involving cells.[26,27,28] One of these areas is food safety, where researchers investigate such things as growth of microorganisms in human food.[27] Producing new knowledge on pathogen adaptation and survival in food could be an important part of reducing incidence rates of pathogen outbreaks caused by contaminated food, which is becoming important as the world's food supply becomes increasingly globalized.[29] In developing countries diarrhea is a major source of infant death, and one of the most common causes is infection by pathogenic *E. coli.*[9]

One of the oldest, and most effective way of curbing microbial growth in foods is through salting.[4,30] By increasing the osmolarity of the environment, salting induces osmotic stress in microorganisms, which greatly reduces growth rates.[2,30] Developing methods capable of modelling cell metabolism during osmotic stress would lead to an increased understanding of the systems regulating osmoadaptation, and might result in more efficient preservation of food.[30] Osmoadaptation is a complex process, involving large changes to cell metabolism. Modelling of this phenomenon can be aided by the development of methods which are able to accurately translate gene expression data into changes in the metabolic networks of cells.

A challenge facing researchers working on gene expression readings today is the lack of 1:1 correlation between gene expression levels and the resulting protein levels.[31] Progress in elucidating the relationship between mRNA and protein levels has progressed at a steady pace, and recently the concept of ribosomal profiling was introduced, where the ribosomal occupancy fraction of mRNAs is profiled across a wide range of mRNAs.[32,33]
Ribosomal profiling is capable of generating a unique “Translation efficiency” parameter to mRNAs – a relative measure of the rate of protein translation from each mRNA.[33] Coupling translation effiency parameters with gene expression readings and knowledge on the mean ratio of protein abundance to mRNA abundance could open the door for methods that predict enzyme concentrations at empirically measured levels, bringing systems biology one step closer to truly representative metabolic models.

This thesis will detail the development of one such method, with the goal of modelling osmotic stress in *E. coli.*

Chapter 2

Theory

This chapter will give an overview of the methods used in the development and subsequent analysis of the "Metabolic Flux Distribution by Translational Efficiency and Enzyme Kinetics" (MUTE) method, which was developed to reach the primary goal of this thesis – understanding osmotic stress in *Escherichia coli* (*E. coli*). The theory should give the reader knowledge about the motivations for understanding osmotic stress in *E. coli*, how metabolic models are constructed and how they are analyzed. It will give an overview of existing methods for integrating biological data with metabolic models, and show why developing the MUTE method was neccessary.

2.1 *Escherichia coli*

E. coli is an enterobacterium which is present in the gastrointestinal tract of humans.[6] *E. coli* has become a massively popular and well understood model organism due to its relatively small, fully sequenced genome of 4.6 Mbp along with a short doubling time of approximately 30 minutes.[34,35] The EcoCyc database, dedicated to knowledge on *E. coli* strain K-12 substrain MG1655 lists 4501 genes, of which 4282 are protein coding.[36] Out of the 4501 genes, 3547 have known functions, representing 78% of the genome.[36]

2.1.1 Pathogenic *E. coli*

Infection by pathogenic *E. coli* manifests mainly in three general clinical symptoms: Enteric/diarrhoeal disease, urinary tract infection and sepsis.[6] A common source of infection for humans is through contaminated food, where pathogenic *E. coli* strains are able to colonize the gastrointestinal

tract and cause disease.[6] Understanding ways of preventing or slowing growth of *E. coli* in food is therefore an important preventative measure that could reduce the incidence rates of *E. coli* infections in humans.
Perhaps the most widely used and oldest preventative measure against pathogen contamination in food is salting, which induces osmotic stress upon micro
-organisms inhabiting food.[4]

2.1.2 Osmotic stress in *E. coli*

Osmotic stress occurs whenever the water activity surrounding a living cell differs from that on the inside.[2] Goverened by the unstoppable force of equilibrium, water will migrate across cell membranes, the direction of which depends on the nature of the inequality, until
equilibrium is achieved.[2]

Because salt is added as a preservative to food, microorganisms living there are typically faced with a hyperosmotic environment, where the water activity of its surroundings is higher than inside the cell.[2] This difference in water activity will over time dehydrate the cell, as water migrates from the cytoplasm to the extracellular environment.[2] To combat this, cells produce osmoprotectants such as glycine-betaine, which help equalize the
water activities inside and outside of the cell.[37,38,3]

Research on osmotic stress in *E. coli* has previously shown that osmotic and oxidative stress is connected, and it is likely that an understanding of one of these systems will increase the understanding of the other.[3,39,40,41] Progress in understanding osmotic stress relies in part on the ability to model the phenomenon *in silico*, using our lists of reactions, genes and metabolites to build mathematical models which represent the metabolisms of microorganisms.[42]

2.2 Reconstruction of genome scale metabolic models

An important aspect of modelling cells' metabolisms is how to represent the individual reactions and molecular species that participate. The most common method is to represent some reaction n as a linear equation, where all metabolites that participate in the reaction have coefficients equal to their stoichiometric coefficients, and all other metabolites have coefficients equal to zero.[43] These coefficients are negative in the case of substrate, and positive for products, representing the destruction or creation of molecular species.[43]

2.2.1 Stoichiometric matrices

Using Michaelis Menten kinetics, reactions catalyzed by an enzyme E can be represented by the following schematic:[44]

$$\mathrm{E} + \mathrm{S} \underset{k_2}{\overset{k_1}{\rightleftharpoons}} \mathrm{ES} \xrightarrow{k_{\mathrm{cat}}} \mathrm{E} + \mathrm{P} \tag{2.1}$$

which is represented by the following differential equation:[44]

$$\nu = \frac{d[P]}{dt} = \frac{V_{\max}[S]}{K_M + [S]} \tag{2.2}$$

Assuming assuming that enzymes are working at saturation and invoking the steady state assumption, i.e. $[S] >> [K_m]$, $V_{\max} = Ek_{\mathrm{cat}}$ allows simplification of equation (2.2):[44]

$$\nu = \frac{d[P]}{dt} = \frac{V_{\max}[S]}{[S]} \implies \nu = \frac{d[P]}{dt} = Ek_{\mathrm{cat}} \tag{2.3}$$

where k_{cat} is the turnover rate of the enzyme E.[44] All reactions in an biochemical network can be represented this way, and assuming mass balance, i.e. $\frac{dP}{dt} = 0$, it is possible to contruct linear stroichiometric matrices where each row represents the mass balance of a metabolite. For a set of reactions, the stoichiometric matrix $\bar{\bar{S}}$ is of the form:

$$\bar{\bar{S}} = \begin{bmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,n} \\ c_{2,1} & c_{2,2} & \cdots & c_{2,n} \\ \vdots & \vdots & & \vdots \\ c_{m,1} & c_{m,2} & \cdots & c_{m,n} \end{bmatrix} \tag{2.4}$$

where $c_{m,n}$ represents the stoichiometric coefficient of metabolite m in reaction n. As an example, consider the five reactions and three metabolites listed below:

$$\longrightarrow \mathrm{A} \tag{2.5}$$

$$\longrightarrow \mathrm{B} \tag{2.6}$$

$$\mathrm{A} + 3\,\mathrm{B} \longrightarrow 2\,\mathrm{C} \tag{2.7}$$

$$2\,\mathrm{A} + \mathrm{B} \longrightarrow 3\,\mathrm{C} \tag{2.8}$$

$$\mathrm{C} \longrightarrow \tag{2.9}$$

The numbering system for the ensuing stoichiometric matrix will be the same as for the equations. For this set of equations, the stoichiometric matrix would be:

$$\bar{\bar{S}} = \begin{bmatrix} 1 & 0 & -1 & -2 & 0 \\ 0 & 1 & -3 & -1 & 0 \\ 0 & 0 & 2 & 3 & -1 \end{bmatrix} \tag{2.10}$$

Looking at the reactions it is evident that A is created in reaction 1 and consumed in reaction 3 and 4. This is represented in the first row stiochiometric matrix, where $S_{1,1} = 1$, $S_{1,3} = -1$ and $S_{1,4} = -2$.

These matrices uncouple biochemical reactions from their enzymes' turnover rates, relying instead on the principle of mass balance, which opens up several possibilities for modelling the biochemical reactions governing a cell. Examples include Extreme Pathways, Elementary mode analysis, Minimal metabolic behaviours, Metabolic Modelling with Enzyme Kinetics and Flux Balance Analysis.[43,45,46,47,48] Metabolic models have been compiled for several organisms, and as of 2009 the number of metabolic models counted 94 for bacteria, 39 for eukaryota and 6 for archaea. Some of the most complete models today describe the metabolism of *E. coli*.[49,50,51]

2.2.2 Metabolic models of *E. coli*

Our detailed understanding of *E. coli* has allowed the creation of metabolic models of the organism, where large parts of the organisms metabolic network is represented by systems of linear equations.[50] The *i*AF1260 reconstruction of *E. coli* metabolism accounts for 1260 Open Reading Frames (ORFs) and 2382 biochemical reactions, making it possible to model aspects of the organisms metabolism with relatively high degrees of accuracy.[50]

Orth and coworkers published an updated genome scale reconstruction of *E. coli* metabolism, *i*JO1366, accounting for 1366 genes, 2251 metabolic reactions and 1136 unique metabolites.[51] The *i*JO1366 model is made in Systems biology Markup Language (SBML), and can be imported by several software suites where a range of methods can use the model as a structure to work on.[52] One of the most popular software suites for metabolic modelling is the Constraints Based Reconstruction and Analysis (COBRA) toolbox for Matlab, in which methods such as Flux Balance Analysis (FBA) can be run to optimize flux distributions in metabolic models.[43,53]

2.3 Linear and quadratic programming for optimization

The field of optimization methods attempts to tackle problems where some optimal distribution of values need to take place, while simultaneously making sure that the constraints of the problem are not violated.[54] Research into optimization has benefited from a large interest in solving complex

logistics problems, and is quickly evolving.[55] While typically associated with engineering and business problems, optimization theory has applications in almost all fields, including metabolic modelling.[43]

Optimization problems are commonly defined by an objective function and a set of constraints.[54] The objective function, as the name implies, represents information about the value of the variables, while constraints represent limits on the range of values variables can take. When optimizing, the goal of the algorithm of choice is to either minimize or maximize the value of the objective function by selecting an optimal combination of values for the variables.[54] The algorithms are able to do this while making sure that the limits imposed by the constraints of the problem are not violated, ensuring that the solution lies within the feasible region of the solution.[54]

2.3.1 Linear programming optimization

Linear programming optimization is applicable to a problem when both the objective function and the constraints of a problem can be expressed as linear functions. The most well known optimization method for linear problems is perhaps the "Simplex algorithm", which was developed by George Dantzig in 1947. The following quote from Dantzig describes the method succinctly:[56]

> "*The simplex procedure is a finite iterative method which deals with problems involving linear inequalities in a manner closely analogous to the solution of linear equations or matrix inversion by Gaussian elimination.*"

The method can be described in the following way:

$$\begin{aligned} \min\ z &= \bar{c}^T\bar{x} \\ \text{subject to} & \\ \bar{\bar{A}}\bar{x} &\leq \bar{b} \\ x &\geq 0 \end{aligned}$$

where $\bar{\bar{A}}$ is the constraint coefficient matrix, $\bar{b}$ is a column vector describing the limits of all constraints in $\bar{\bar{A}}$ and $\bar{c}$ is a row vector of objective function coefficients. Following the optimization step iterations, a vector containing optimal values of $\bar{x}$ is returned.[56]

The simplex algorithm moves from vertex to vertex along the n–dimensional surface of the feasible solution space.[56] If a point is found, where moving from that point in any direction decreases the value of the objective function, that point represents an optimal combination of values for the variables.[56]

Example of Simplex agorithm

Consider the following linear problem:

$$\max\ z = 5x_1 + 2x_2 \tag{2.11}$$

$$\text{subject to}$$

$$4x_1 - x_2 \geq 0 \tag{2.12}$$

$$-x_1 + 2x_2 \leq 7 \tag{2.13}$$

$$x_1 + 2x_2 \leq 13 \tag{2.14}$$

$$x_1 + x_2 \leq 9 \tag{2.15}$$

$$2x_1 - x_2 \geq 9 \tag{2.16}$$

$$x_1 - 2x_2 \geq 3 \tag{2.17}$$

$$x_1, x_2 \geq 0 \tag{2.18}$$

In order to solve this, the Simplex algorithm moves from vertex to vertex along the surface defined by the constraints of the problem. At each vertex, the value of the objective function in that point is checked, and if it is greater than the current value, the Simplex algorithm moves to that point. If the objective function value in the next vertex is lower than the current, the global optimum has been found, and the algorithm is finished.

This is caused by the nature of linear problems, which are convex in nature.

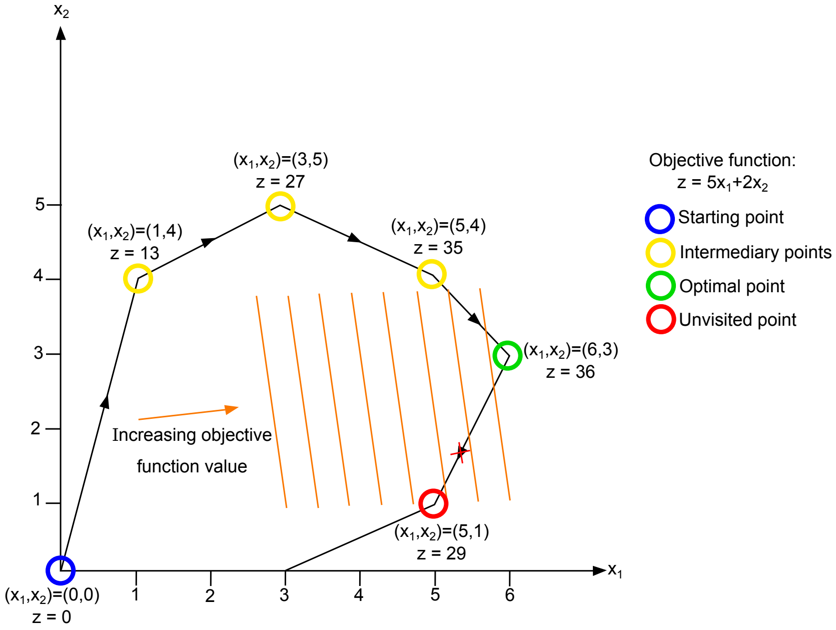

Figure 2.1: Illustration of the simplex algorithm.

In convex systems, local and global minima or maxima are equivalent.[57] Figure 2.1 shows a graphical representation of the Simplex algorithm solving the linear problem defined in equations (2.11) through (2.18).

In Figure 2.1, the algorithm starts in $(x_1, x_2) = (0, 0)$. It then travels along the first constraint, arriving in the second vertex, $(x_1, x_2) = (1, 4)$, where the objective function value improves. This is repeated until the algorithm reaches the optimal point $(x_1, x_2) = (6, 3)$. The next vertex objective function value, $z = 29$ is lower than the optimal point, $z = 36$, and the algorithm is finished. Note that all vertices represent integer points merely for convenience, and that the Simplex algorithm is not restricted to integer programming problems.[58]

2.3.2 Quadratic programming optimization

In quadratic optimization, both the objective function and the constraints can have quadratic terms.[54] Quadratic problems are more complex than their linear counterparts due to the inclusion of quadratic variables, and therefore need separate algorithms for solving.[54] A quadratic optimization problem will typically be of the form:

$$\begin{aligned} \min z &= \bar{c}^T\bar{x} + \bar{x}^T\bar{\bar{Q}}\bar{x} \\ \text{subject to}& \\ \bar{\bar{A}}\bar{x} &\leq b \\ x &\geq 0 \end{aligned}$$

where $\bar{\bar{A}}$ is the constraint coefficient matrix, $\bar{\bar{Q}}$ is the quadratic objective coefficient matrix, $\bar{x}$ is a vector of variables and $\bar{c}$ is the linear objective coefficient row vector[54]. Compared to linear programming optimization, where the objective function is convex, quadratic programming problems have concave objective functions. Checking for local maxima or minima for concave functions has been shown to be NP-hard – a class of problems with no known algorithms for exact solutions.[59] Optimizing quadratic programming problems therefore requires algorithms, such as the Barrier algorithm, which make use of certain heuristics in order to converge to optimal solutions.[60]

Quadratic programming optimization can be used for problems such as minimization of distance between some number and a variable, for instance the minimization of distance between empirical data and predictions.[61]

2.4 Flux Balance Analysis

Flux Balance Analysis (FBA) is one of the constraint based approaches to metabolic modelling.[43] Expanding on the metabolic network defined by the stoichiometric matrix $\bar{\bar{S}}$, FBA adds linear constraints that shape the solution space of the metabolic model.[43]

2.4.1 Linear constraints

Linear constraints make it possible to constrain the solution space of a model by imposing sets of strict (in)equalities on the system. Single variables, or the sum of several can be prevented from, or forced to, take on values defined by some numerical limit.[43] Any constraint is of one of the following forms:

$$s_1x_1 + s_2x_2 + ... + s_nx_n \leq b \tag{2.19}$$

$$s_1x_1 + s_2x_2 + ... + s_nx_n \geq b \tag{2.20}$$

$$s_1x_1 + s_2x_2 + ... + s_nx_n = b \tag{2.21}$$

Using various combinations of these three types of constraints, it is possible to ensure that models behave in a way that respects *a priori* knowledge, resulting in more accurate and applicable predictions. An important part of FBA is the following set of restrictions, represented here in matrix form:[43]

$$\bar{\bar{S}} \cdot \bar{\nu} = \bar{0} \tag{2.22}$$

This set of constraints restricts the model in such a way that the sum of fluxes for any given metabolite is zero. Implementing this prevents accumulation of metabolites in biochemical dead ends, and ensures that reaction fluxes flow through the model according to the logic defined in $\bar{\bar{S}}$.[43] Other constraints limit the lower or upper bound of reaction fluxes, keeping them at realistic, empirically validated levels. Linear constraints are also used to define "virtual growth medium" of the metabolic models, limiting the nutrient availability of the cells. These constraints change the shape of the solution space of the model, as can be seen in Figure 2.2.[43] Each edge in the cone-like structure of Figure 2.2 is defined by some constraint, and together they form a subspace of the unconstrained solution space.

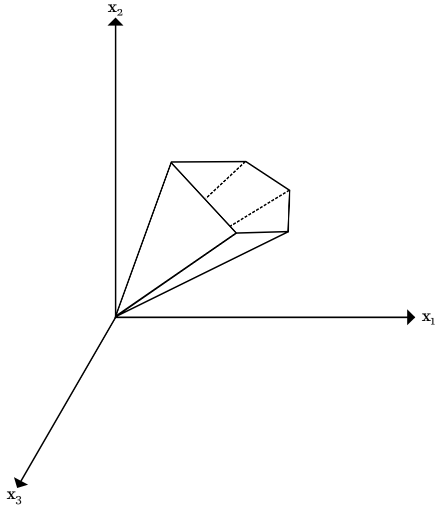

Figure 2.2: The solution space of a 3-dimensional linear model.

Finally, one last piece is required for FBA, and that is an objective function which drives flux through the reactions.

2.4.2 Linear objective functions

The objective function, as the name implies, defines the objective of the model, or in the case of a metabolic model, the objective of the cell.[43] The function is commonly linear, and associates some value c with all variables.[43] A typical objective function is on the form:

$$\max z = \bar{c}^T \cdot \bar{x} \tag{2.23}$$

where $\bar{c}$ is a vector containing the value associated with all variables in $\bar{x}$. The inclusion of "max" in front of the function is an instruction to the optimizer that the value of the objective function is to be maximized. The most common objective function in FBA is the accumulation of biomass, which will drive the optimization towards a distribution of reaction fluxes which maximizes the growth rate of the modelled cell.[43]

The optimization step of FBA is handeled by external optimizers such as Gurobi[62], GLPK[63] or CPLEX[64], and returns a vector of predicted reaction fluxes for all reactions in the model, along with the predicted growth rate of the modelled cell.[43] FBA has been a popular choice for researchers investigating such phenomenon as the effect of gene knockouts on metabolic networks, and has been of great use in disciplines such as metabolic engineering.[65,66] The downside to FBA is that it is unable to accomodate OMICS data in the form of gene expression, proteomics– or enzyme kinetics data. Modelling complex and poorly undestood systems such as osmotic stress relies on these kinds of data to give clues as to what is happening, and so developing methods which incorporate OMICS data is an area of high activity.[67]

2.5 High throughput measurements of biological data

The massive generation of data which has been made possible by relatively recent advances in molecular biology has allowed us to gain detailed insights into the inner workings of cells and the flow of information which regulates life. The OMICS term encompasses many data types, among them data on gene expression, protein concentrations, metabolite concentrations and reaction fluxes. Finding ways of integrating various types of OMICS data has been the goal of many groups working on systems biology, and the number of methods incorporating these kinds of data are rapidly increasing.[67]

2.5.1 Proteomics

The conventional proteomics methods make use of 2–dimensional gel separation in order to generate maps of protein abundances.[68] Proteins are applied to wells in a polyacrylamide (PAGE) gel, which is then submerged in a solution with a pH gradient. Subjecting the system to an electrical current makes the proteins migrate along the electrical field, until they are at a point in the pH gradient where their natural charge is neutralized, stopping that particular protein's migration.[68] After some time, all proteins in the applied sample will have separated along one dimension based on their isolelectric point.[68]

After separating the proteins based on isoelectric points, a solution containing charged detergents, most commonly sodium dodecyl sulfate (SDS) is added to the protein samples in order to linearize them an impart negative charge to the proteins.[68] SDS typically distributes evenly along proteins, giving all proteins an approximately equal charge to mass ratio.[68] A new electric current is applied to the PAGE gel, orthogonal to the previous one. All proteins now migrate along the new electric field, and separate in the second dimension based on mass, as larger proteins migrate more slowly than smaller ones.[68]

After completing the two dimensional separation of proteins based on isoelectric points and mass, the proteins can be transferred to other surfaces, such as nitrocellulose membranes, where they can be visualized and characterized more easily.[69]

Two-dimensional separation of proteins on gels is losing ground to modern methods such as High Pressure Liquid Chromatography (HPLC) and affinity purification, which are faster and more accurate, but it is important to recognize the role of 2–dimensional gel separation has played in proteomics history.[70,71,72]

Isolating proteins has made it possible to determine several of their characteristics, such as enzyme turnover rates and post-translational modifications. Protein data is becoming more and more important, and methods such as Metabolic Modeling with Enzyme Kinetics (MOMENT) utilize information on protein weights and enzyme turnover rates to make flux- and growth rate predictions from metabolic models.[48]

2.5.2 Gene expression microarrays

Technologies such as gene expression microarrays have made it possible to detect and measure mRNA levels from the complete set of genes in a genome simultaneously, and investigating the mRNA levels from specific genes under different environmental conditions has allowed the elucidation of the functions of many genes.[73]

Microarrays rely on the principle of hybridization of nucleotides.[74] This hybridization facilitates the design of oligonucleotide probes, which are able to hybridize with complementary sequences of DNA/RNA.[74] The oligonucleotide probes are labelled with fluorescent dyes, such as Cy3 and Cy5, which emit characteristic spectra of light when excited by laser light.[74,75]

A microarray is produced by printing a set of probes into wells in a glass plate. The set of probes is selected with a specific experiment in mind, making sure that each mRNA of interest has a complementary probe attached to the surface of some well on the array. A sample of purified mRNA is amplified by reverse transcription, producing a corresponding set of cDNA sequences complimentary to the set of mRNAs.[74,75]

The cDNA sample is applied to the microarray, where the cDNA sequences hybridize with complimentary probes. Excitation of the fluorescent dyes in the probes by laser produces a signal for each oligonucleotide probe which depends on emitted light intensity. These signals are analyzed and used to measure the original level of mRNA in the experimental sample, represented by some probe on the array.[74,75]

Gene expression microarrays are a valuable tool for eludicating gene functions, and methods such as Metabolic Adjustment by Differential Expression make use of gene expression data to predict changes in the metabolism of cells.[25]

2.6 Metabolic Adjustment by Differential Expression (MADE)

MADE was developed as a way of integrating high-throughput expression data with metabolic models.[25] MADE takes as input a set of gene expression measurements which were taken
different environmental/cellular conditions. At least one transition between these conditions is defined by MADE, describing the fold changes in expression levels of the measured genes between the two conditions. The method also relies on the *p–values* associated with these transitions, commonly found by performing a *t-test* on the data series.[25] Armed with fold changes and p–values, MADE

proceeds by grouping each gene's expression change in each tranition into one of three sets; increasing, decreasing or constant. An objective function for each transition is defined as the weighted sum:[25]

$$
\begin{aligned}
f_{i\to i+1}(x) = &\sum_{x\in I} w(p_{x_{i\to i+1}})(x_{i+1} - x_i) && (2.24)\\
&+\sum_{x\in D} w(p_{x_{i\to i+1}})(x_i - x_{i+1}) && (2.25)\\
&-\sum_{x\in C} w(p_{x_{i\to i+1}})\Delta_{x_i,x_{i+1}} && (2.26)
\end{aligned}
$$

where

$$
\Delta_{x_i,x_{i+1}} = \begin{cases} 0, & \text{if } x_i = x_{i+1} \\ 1, & \text{if } x_i \neq x_{i+1} \end{cases}
$$

and $w(p_{x_{i\to i+1}})$ is the weighting function of the *p–value* associated with the transition, typically $-log(p)$. The algorithm then optimizes the sum of all weighted transition sums:[25]

$$
\max \sum_{i=1}^{n-1} f_{i\to i+1}(x) \tag{2.27}
$$

Resulting in an optimal set of binary states for each gene for each condition. Genes whose binary state is set to 0 are predicted by MADE to be inactive in that condition, while genes whose state is 1 are allowed to carry flux.[25]

2.6.1 Ilustrative example of MADE

As an example, consider the following case. The imaginary organism *Nanococcus minimalus* has three genes. A researcher looking to investigate *N. minimalus* sets up a series of experiments where the mRNA levels of each of the three genes is measured in three different conditions. Following the experiments, after analyzing the data, the researcher is left with Table 2.1.

Table 2.1: Example of MADE input data.

	Fold change 1 → 2	p–value	Fold change 2 → 3	p–value
gene 1	2.0	0.001	1.2	0.050
gene 2	0.5	0.003	0.1	0.001
gene 3	0.0	0.001	0.0	0.001

```
MADE: Metabolic Adjustment by Differential Expression
-----------------------------------------------------


...

Gene counts:
           |  Increasing      Decreasing      Constant
Transition |  Fit / data      Fit / Data    Fit / Data
 1 -> 2    |    1 /    1        0 /    1      1 /    1
 2 -> 3    |    0 /    1        1 /    1      1 /    1

Total match:  4 / 6 (66.7%)
```

Figure 2.3: Example of MADE output in Matlab.

There are two transitions between the three conditions, each associated with some p–value. Feeding the data contained in table 2.1 to the MADE algorithm, along with a metabolic model of *N. minimalus* produces the following edited output from Matlab:

The MADE algorithm has adjusted the gene states in the three conditions to match the experimental data from Table 2.1. The expression changes in transitions are prioritized according to the weighting function previously described, where lower p-values are prioritized over higher ones. For this example, the gene state matrix produced by MADE is shown in Table 2.2.

Table 2.2: Gene states predicted by MADE.

Gene	Condition 1	Condition 2	Condition 3
Gene 1	0	1	1
Gene 2	1	1	0
Gene 3	1	1	1

The reactions belonging to the genes in Table 2.2 would have their flux bounds adjusted according to the gene states, where “0” would constrain the reaction to carry zero flux, while “1” would allow normal flux to flow through the reaction.

2.6.2 Strengths and weaknesses of MADE

In the example shown in section 2.6.1, some weaknesses of MADE are made apparent. By utilizing binary variables for gene states, any consecutive increase in expression over multiple conditions will only result in one gene state change. This has the potential of misinterpreting the state of genes, forcing expressed genes in one or more conditions to the off state, because of a significant increase in a later transition. This leads to a "coarse" mode of action for MADE, where big changes such as complete metabolic shifts are represented, while smaller changes that merely alter the flux rates between conditions are lost.[25] Another weakness is the incompatibility of MADE models, which are in the TIGER or ELF format, with many COBRA Toolbox methods, making the analysis of the models challenging.[25,76]

The strength of the MADE method is its ability to represent transitions between different environmental conditions. By optimizing the on/off state of genes over a set of environmental transitions, MADE is able to identify trends in time series measurements, that can help elucidate adaptation mechanisms of cells in various states of environmental stress.[25]

2.7 Metabolic Modelling with Enzyme Kinetics (MOMENT)

The MOMENT method takes an unconventional approach to predicting growth rate and flux distributions of organisms. Whereas some methods incorporate OMICS data in order to increase their predictive accuracy, MOMENT makes use of knowledge about the kinetic parameters of the enzymes which catalyze metabolic reactions, as well as knowledge on the mass composition of cells.[48]

Using enzyme kinetics databases such as BRENDA and SABIO-RK, enzyme turnover rates are collected for all enzymes contained in the model.[77,78] If some enzyme's turnover rate is not available for the organism in question, turnover rates from closely related organisms are used as a substitute, and if this fails, the mean of all enzyme turnover rates are used instead.[48]

The chemical composition of cells is used to set an upper limit on enzyme mass in the cell, and combined with the molecular weights of all enzymes in the model this can effectively constrain the sum of metabolic fluxes. The product of an enzyme i's concentration, denoted g_i, and its turnover rate, denoted kcat$_i$ acts as an upper constraint on the allowed flux through some reaction associated with this enzyme.[48]

MOMENT formulates the following quadratic programming problem:

$$\max \textit{(biomass production)} \tag{2.28}$$

$$\text{subject to} \tag{2.29}$$

$$\bar{\bar{S}} \cdot \bar{\nu} = 0 \tag{2.30}$$

$$v_j \leq \begin{cases} k_{\text{cat}}^j \cdot g_i, & \text{if condition 1} \\ k_{\text{cat}}^j \cdot (g_a + g_b), & \text{if condition 2} \\ k_{\text{cat}}^j \cdot \min(g_a, g_b), & \text{if condition 3} \end{cases} \tag{2.31}$$

$$\sum g_i \cdot MW_i \leq C \left[\frac{g_{\text{protein}}}{g_{DW}} \right] \tag{2.32}$$

where $\bar{\bar{S}}$ is the stoichiometric matrix, $\bar{\nu}$ is a vector of reaction fluxes, k_{cat}^j denotes the turnover rate of reaction j, MW_i denotes the molecular weight of enzyme i, g_i is the predicted amount of enzyme i in the cell, g_{protein} is the total weight of proteins (assumed to be 56% of *E. coli* dry weight mass) and C denotes the fraction of proteins dedicated to metabolic enzymes.[48] Condition 1 refers to a reaction catalyzed by a single enzyme i, condition 2 to a reaction catalyzed by two isozymes a OR b and condition 3 to a reaction catalyzed by an enzyme complex made up of proteins a AND b.

MOMENT is in many ways a different take on the Flux Balance Analysis with Molecular Crowding (FBAwMC) method, in which molecular volume is used in place of molecular weight.[79] Where MOMENT imposes constraints on total enzyme mass, FBAwMC imposes constraints on total enzyme volumes. Due to the low amount of information on the volumes of enzymes, compared to their mass, the MOMENT method is at an advantage in its accuracy, and by extension its predictions.[79,48]

2.8 Metabolic Flux Distribution by Translational Efficiency and Enzyme Kinetics (MUTE)

MUTE was developed as a way of bridging the gap between gene expression data, protein concentrations and metabolic flux. One challenging aspect of gene expression measurements is that mRNA levels only weakly correlate with reaction fluxes. One of the reasons for the lack of 1:1 correlation is the system translating mRNA to proteins; ribosomes. Since reaction flux is ultimately decided by enzyme concentrations and their turnover rates, it is crucial that an understanding of the translation of mRNA to proteins is found, so that it is possible generate to more accurate predictions of enzyme levels and, by extension, reaction fluxes.[80]

The development of MUTE was inspired in part by the MOMENT method, which used enzyme turnover rates and total enzyme mass limits to predict growth rates from metabolic models.[48] The implementation of MOMENT, in which upper bounds on reaction flux are proportional to specific enzyme levels, makes is necessary to optimize two sets of variables: reaction fluxes and the enzyme concentrations. While MOMENT's approach to predicting growth rates is clever, it is unable to accomodate OMICS data such as gene expression readings, and so its applications are limited.[48]

MUTE makes use of gene expression data by coupling each mRNA to a unique translational efficiency parameter, based on experimental measurements.[33] The translational efficiency parameter gives a relative measure of how many proteins each mRNA produces, and the product of the mean number of proteins pr mRNA, a gene's mRNA level and its corresponding translational efficiency parameter should give a good prediction of the resulting enzyme's concentration.[33] The most ambitious part of MUTE is perhaps its strong reliance on empirical values for all calculations without the use of "fudge factors" to better scale predictions to realistic levels.

A limit on the total enzyme mass of a cell is enforced through a linear constraint, in order to avoid scenarios where cells have unrealistically high levels of enzymes.

2.8.1 Problem formulation

MUTE can be separated into three parts:

1. The quadratic programming problem formulation which consists of a quadratic objective function and a set of linear constraints:

$$\min z = \sum_{i=1}^{n} \sqrt{((\mathrm{mRNA}_i \cdot T_i \cdot \kappa) - x_i)^2} \tag{2.33}$$

subject to

$$\sum_{i=1}^{n} \mathrm{MW}_i \cdot x_i \leq \frac{M_{\mathrm{cell}} \cdot F_{\mathrm{enz}} - M_{\mathrm{unaccounted}}}{V_{\mathrm{cell}}} \tag{2.34}$$

$$x_i \leq C_{\max}, \forall i \in \{1, ..., n\} \tag{2.35}$$

$$x_i \geq 0, \forall i \in \{1, ..., n\} \tag{2.36}$$

2. Total enzyme concentrations for a reaction R are calculated as follows:

$$E_R = \sum_{i \in S_R} s_i + \sum_{\substack{C_{R,j} \in C_R \\ c \in C_{R,j}}} \min(c_{j,s}, c_{j,s+1}, ..., c_{j,n}) \tag{2.37}$$

3. And finally, upper flux constraints on reactions are set according to:

$$\nu_R \leq E_R \cdot k_{\mathrm{cat}}^R \tag{2.38}$$

where

- mRNA_i is the mRNA copy number of gene i
- T_i is the translation efficiency of mRNA_i
- $\kappa = \frac{P_{\mathrm{med}}}{N_A \cdot V_{\mathrm{cell}}}$, P_{med} is the median copy number for proteins for the cell, N_A is Avogadro's number and V_{cell} is the volume of the cell in liters
- x_i is the predicted protein concentration of gene i
- F_{enz} is the fraction of cell mass dedicated to metabolic enzymes
- $M_{\mathrm{unaccounted}}$ represents the mass of metabolic enzymes which is unaccounted for by the gene expression data
- $C_{\max}$ is the maximum concentration of any given enzyme. In MUTE's current implementation $C_{\max}$ is set to 2 μmol, or approximately 8000 molecules per cell.

- E_R is the total enzyme concentration for reaction R
- S_R is the set of single enzymes catalyzing reaction R
- s_i is a single enzyme belonging to the set S_R
- C_R is the set of enzyme complexes catalyzing reaction R
- $C_{R,j}$ is an enzyme complex belonging to the set C_R
- $c_{j,s}$ is a subunit of enzyme complex $C_{R,j}$
- ν_R is the flux of reaction R
- k_{cat}^R is the enzyme turnover rate for reaction R

In order to properly assign upper flux bounds to reactions based on their respective enzymes' concentrations, every reaction is represented to a binary AND/OR tree, like the one shown in Figure 2.4. For each reaction, the concentrations of each isozyme is summed, while enzyme complexes are assigned a concentration equal to the minimum concentration of any of its subunits. The enzyme concentration for the whole reaction, shown as the green node in Figure 2.4 is the sum of the concentrations of all isozymes (shown as blue nodes), including any eventual complexes (shown as red nodes).

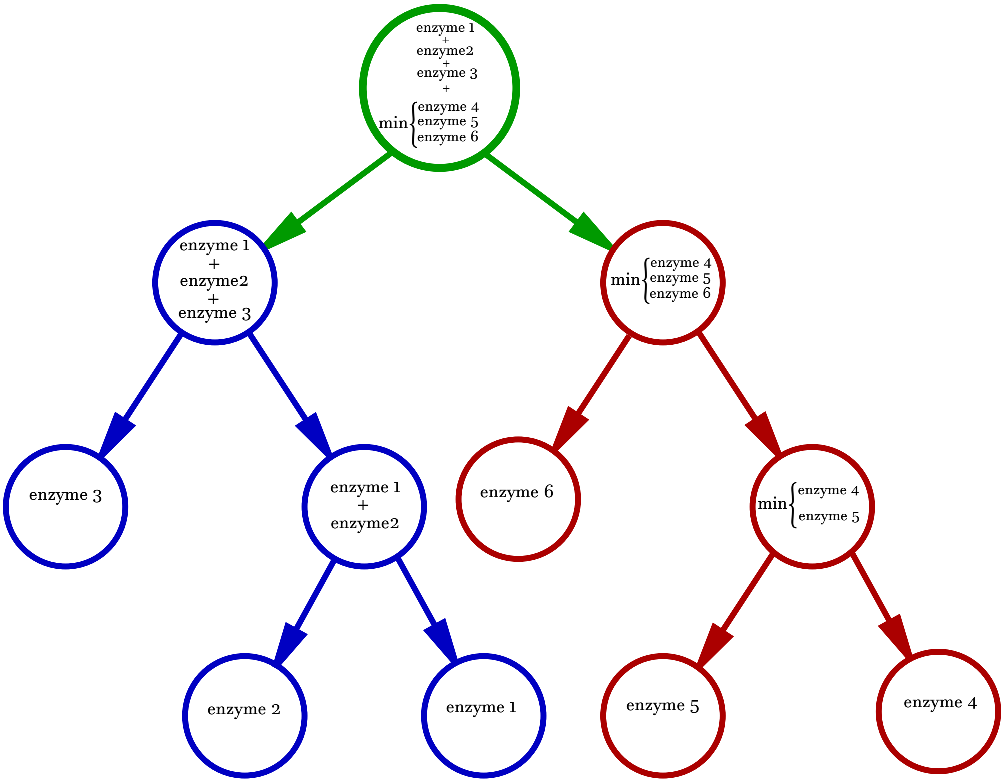
enzyme 1 + enzyme2 + enzyme 3 + min{enzyme 4, enzyme 5, enzyme 6}
enzyme 1 + enzyme2 + enzyme 3
min{enzyme 4, enzyme 5, enzyme 6}
enzyme 3
enzyme 1 + enzyme2
enzyme 6
min{enzyme 4, enzyme 5}
enzyme 2
enzyme 1
enzyme 5
enzyme 4

Figure 2.4: Binary search tree for reaction (green) catalyzed by three isozymes (blue) and an enzyme complex (red).

In the end, MUTE returns a modified SBML model object which has placed new limits on the upper bounds of reaction fluxes, which can be further analyzed by all other methods using SBML model objects as their basis structure, such as sampling methods.[52,53]

2.9 Data sampling

When dealing with many–dimensional optimization problems it can be difficult to get a sense of the distribution of allowed values for the variables involved. A 2–dimensional problem with linear constraints on some interval $x_{\min} \leq x \leq x_{\max}$, $y_{\min} \leq y \leq y_{\max}$ can easily be visualized as some square, 2–dimensional surface. However, when visualizing thousands of variables the brain struggles to keep up. One way of investigating these distributions of such problems is through sampling.[81]

One approach to sampling is to enclose our feasible region S in a larger, simpler region R, which is then sampled. Each sampled point is checked for membership to the feasible region S, and rejected if it is not a member. With enough such samples, the feasible region S can be effectively mapped. This is commonly referred to as rejection methods.[82]

2.9.1 Unbiased random sampling of the solution space of metabolic models

Classical metabolic modelling methods such as Flux Balance Analysis will return a flux vector solution that is merely one of many equally good solutions. The selected solution will be one of the corner points of the hypersphere which is defined by the constraints of the problem, i.e. where several constraints intersect, Figure 2.2. Along the isocline of the optimal objective function value, many alternative flux distributions exist, however only one is presented to the user. This can result in the drawing of untrue conclusions from the predictions made by the optimization method, and care must be taken to avoid this. One method for achieving this is sampling the nullspace of the model.[81]

Using sampling methods on metabolic models makes the user able to generate a range of allowed fluxes for all reactions. It is important that enough sample points are recorded, in order to ensure that a representative image of the null space is achieved.[81]

2.9.2 Hit and Run sampling

Hit and Run sampling is a fast and popular method for sampling convex spaces.[83] In its essence, the Hit and Run sampler can be thought of as an arrow reflecting off of edges representing the constraints defining the solution space of the model, Figure 2.5:

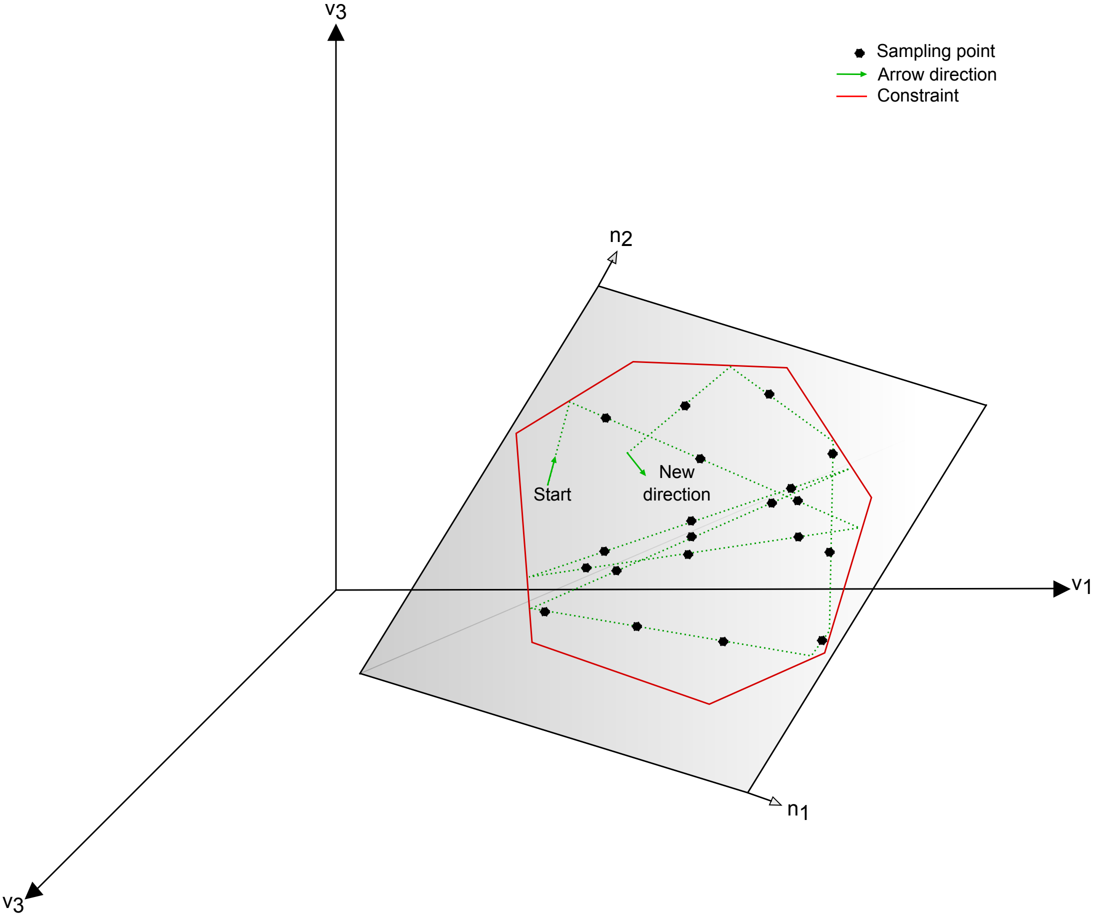

Figure 2.5: Hit and run sampling of a 3-dimensional solution space.

The arrow is shot in a random direction from within the solution space of the model. At some point, the arrow hits an edge/constraint, and is reflected at an outgoing angle equal to the incoming angle. The reflection changes the direction of the arrow, sending it on its way until it encounters another edge/constraint, where it is once again reflected. At even intervals among the arrows path through the solution space, the coordinates of the arrow are recorded. These are the sampling points. After some set number of points have been sampled the arrow is stopped, and a new random direction is chosen from the arrows current point. This whole procedure continues until some predefined number of points have been sampled.[83]

In order to ameliorate numerical instability, the solution space of the problem is transformed from an n–dimensional to a 2–dimensional space. After the sampling is done, all points are transformed back to their original dimensionality.[83] Figure 2.5 gives a graphical illustration of sampling for one random direction in a Hit and Run sampler. The solution space is in this case 3-dimensional (denoted as v), which is transformed to a 2–dimensional (denoted as n) surface where the sampling takes place.

2.10 Principal Component Analysis

When working with many–dimensional problems, findings ways of analyzing and visualizing the data can be challenging. Section 2.9 describes how random sampling of solution spaces can help the investigation of a problem, but the sample points will in most cases still represent some n–dimensional space, where n is arbitrarily large. Principal component analysis (PCA) is able to represented multi–dimensional data in a reduced dimension space.[84]

In a many–dimensional data set, PCA creates linear combinations of variables, called principal components, such that the points that make up the data set are maximally separated when mapped onto the new "principal component dimension".[84] The first principal component displays the largest amount of variance, and the components proceeding it represent the largest amount of variance while orthogonal to all previous principal components.[84] By mapping each point onto the first two principal components, it is possible to transform an n–dimensional data set into a 2–dimensional one, which is easily visualized.[84] The visualization, when paired with categorization of the points, can reveal subtle differences between sets of measurements. Looking at the scores (weights) of the variables in each component can help identify variables which contribute significantly to the phenomenon being investigated.[84]

Chapter 3

Methods

All methods were used within the COBRA Toolbox framework for Matlab R2013a.

3.1 Generation of MADE models

MADE models were generated using the TIGER Toolbox for Matlab. Gene expression data sampled from *E. coli* grown in glucose minimal media with varying NaCl concentrations (2.0 %, 3.5 %, 4.5 %, 5.0 % and 5.5 % respectively)[3], and otherwise identical were used as input to the MADE method. The gene expression data used did not have p-values attached to each reading, and so an alternative approach was used where the gene expression data set for each salt concentration was fit to a normal distribution. This normal distribution was then used to assign p-values to all readings, allowing their use with MADE.

3.2 Visualization of MADE flux predictions

The MADE models were analyzed using the TIGER Toolbox' Flux Balance Analysis (FBA) method. The flux distributions predicted by FBA were visualized by overlaying the reaction fluxes onto metabolic maps of *E. coli* central metabolism.

3.3 Generation of MUTE models

The gene expression data used as input for MUTE was the same as for MADE. Turnover numbers for *E. coli* enzymes, reported by Adadi et. al.[48]

were used for K_{cat}-values. Molecular weights for *E. coli* enzymes were downloaded from the EcoGene database[85]. The translation efficiency parameters used were reported by Li et. al.[33]

The gene expression-, translation efficiency-, enzyme turnover rate-, and molecular weight data were filtered against each other, so that only genes where data from all four sources existed were included. The resulting data set, in addition to the *i*JO1366 *E. coli* metabolic reconstruction was fed to the MUTE method. Each data set fed to MUTE resulted in a separate SBML model structure, with upper bounds placed on all reactions where data from all five sources (reaction exists in reconstruction, gene expression values, enzyme molecular weights, enzyme turnover numbers, translation efficiencies) were present.

An alternative data filtering method was applied to the input data, where genes with no enzyme turnover rate- or translation efficiency data had these parameters set to the average of the enzyme turnover rate or the translation efficiency, respectively. Five alternative MUTE models were generated, where the number of reactions with changed upper flux bounds was much greater than for the previous approach.

3.4 Visualization of MUTE constraints

The upper flux bounds imposed by MUTE were visualized by overlaying the flux bounds onto a metabolic map of the *i*JO1366 model. The metabolic map was downloaded from the BiGG Database[86]. These metabolic maps were made so that the reactions (edges) connecting the metabolites were color coded according to flux size and directionality, making them intuitive to interpret. Reactions through which there was no flux were colored white, to enhance the visibility of the active reactions. These upper flux bound visualizations were done for the five models produced with “strict data filtering” and the five models without “strict data filtering”, and can be found in the appendix.

3.5 Analysis of MUTE models by unbiased random sampling

Flux variability analysis (FVA) was performed on all five MUTE-models, in addition to the original *i*JO1366 model. FVA tested all reactions contained in the models, and the lower limit for optimal growth rate was set to 100 % of maximum growth rate.

The lower and upper flux limits predicted by FVA for all reactions were compared between all “strict data filtering” models, and those with equal upper and lower limits were filtered out. The solution space of these reactions was sampled using the optGpSampler[87] on all five MUTE models, sampling 10,000 points for each reaction with a step length of 2 between each sample point. All models had their growth rates locked to the optimum value, as predicted by FBA, before the sampling. The data sampling procedure was repeated for the MUTE models generated without “strict data filtering”.

The sampled points from both data filtering regimes were analyzed separately using Principal Component Analysis (PCA). All points from the samples were normalized and transformed into a point in the two-dimensional space defined by the two first principal components. The transformed points were visualized in a scatter-plot. Each point in the scatter plot was grouped into one of the five salt concentrations, based on which MUTE model it belonged to.

After running PCA on the sample data sets, the principal components were investigated for interesting reactions. Every reaction in the model was associated with some weight, and the reactions who’s weights had a magnitude of 0.1 or more were checked for involvement in processes such as osmotic- or oxidative stress.

A different approach to identifying important reactions was conducted. Reactions where the flux within each sample had low variance, but where the mean flux of those reactions varied greatly between each sample were identified and investigated. This was done for MUTE models both with and without strict data filtering.

Chapter 4

Results and discussion

Introduction to reading flux maps

Maps of metabolic fluxes represent a large portion of the results in this chapter. This section will give a brief overview of how to read them.

Flux maps

In this thesis' metabolic flux maps, red colors represent positive fluxes while green fluxes represent negative fluxes. Color intensities scale linearly with flux strength, such that the most intense colors represent the largest fluxes.

4.1 Using MADE to predict flux changes

Owing to the lack of compatibility between the model structures returned by MADE and many COBRA Toolbox methods, the MADE models were unable to undergo sampling. All flux predictions for MADE are therefore those produced by FBA.

The flux distribution in the central metabolism of *E. coli* for the five MADE models, as predicted by FBA is shown in Figures 4.1 to 4.5.

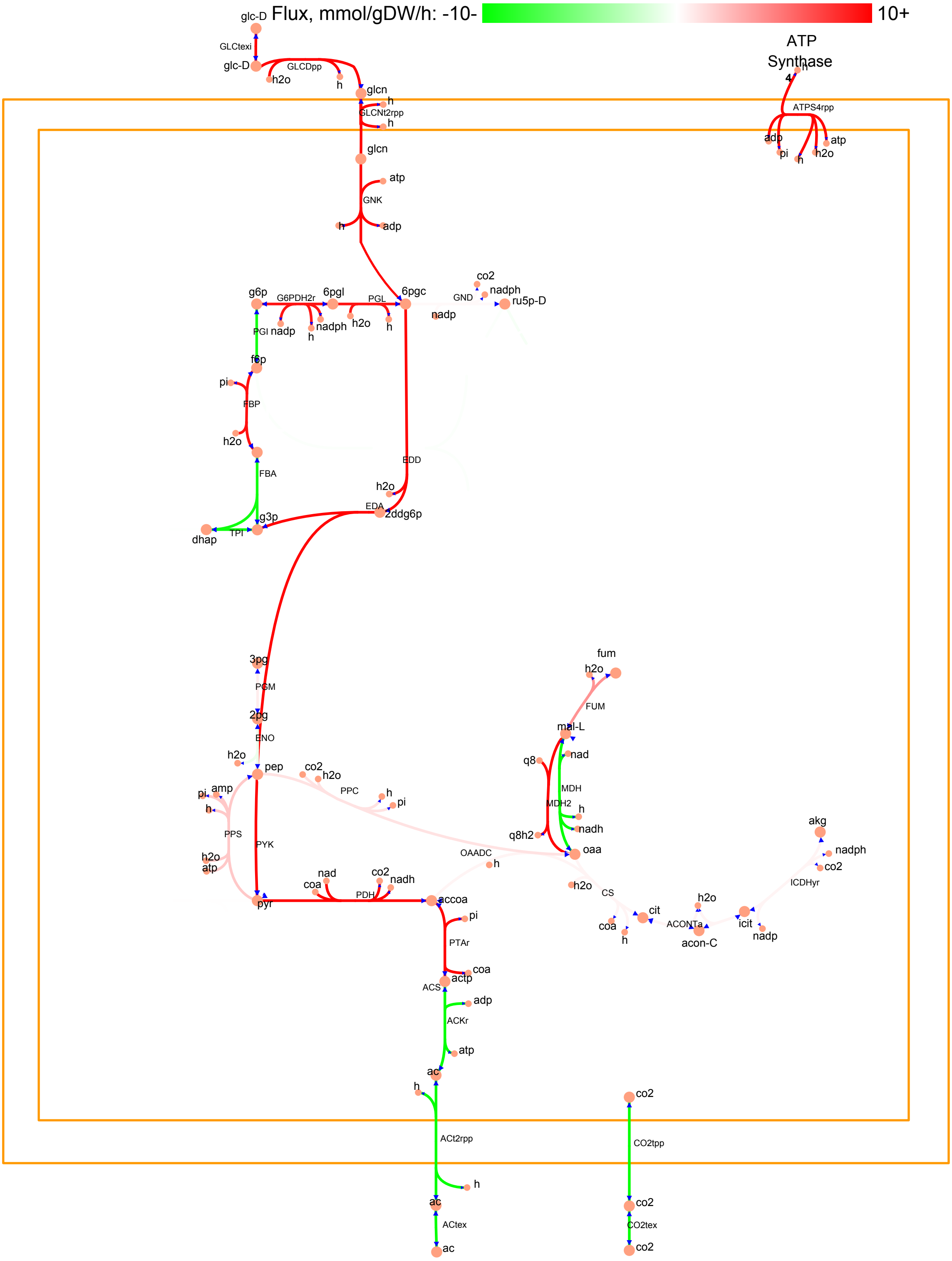

Figure 4.1: MADE FBA core metabolism flux predictions for 2.0 % salt.

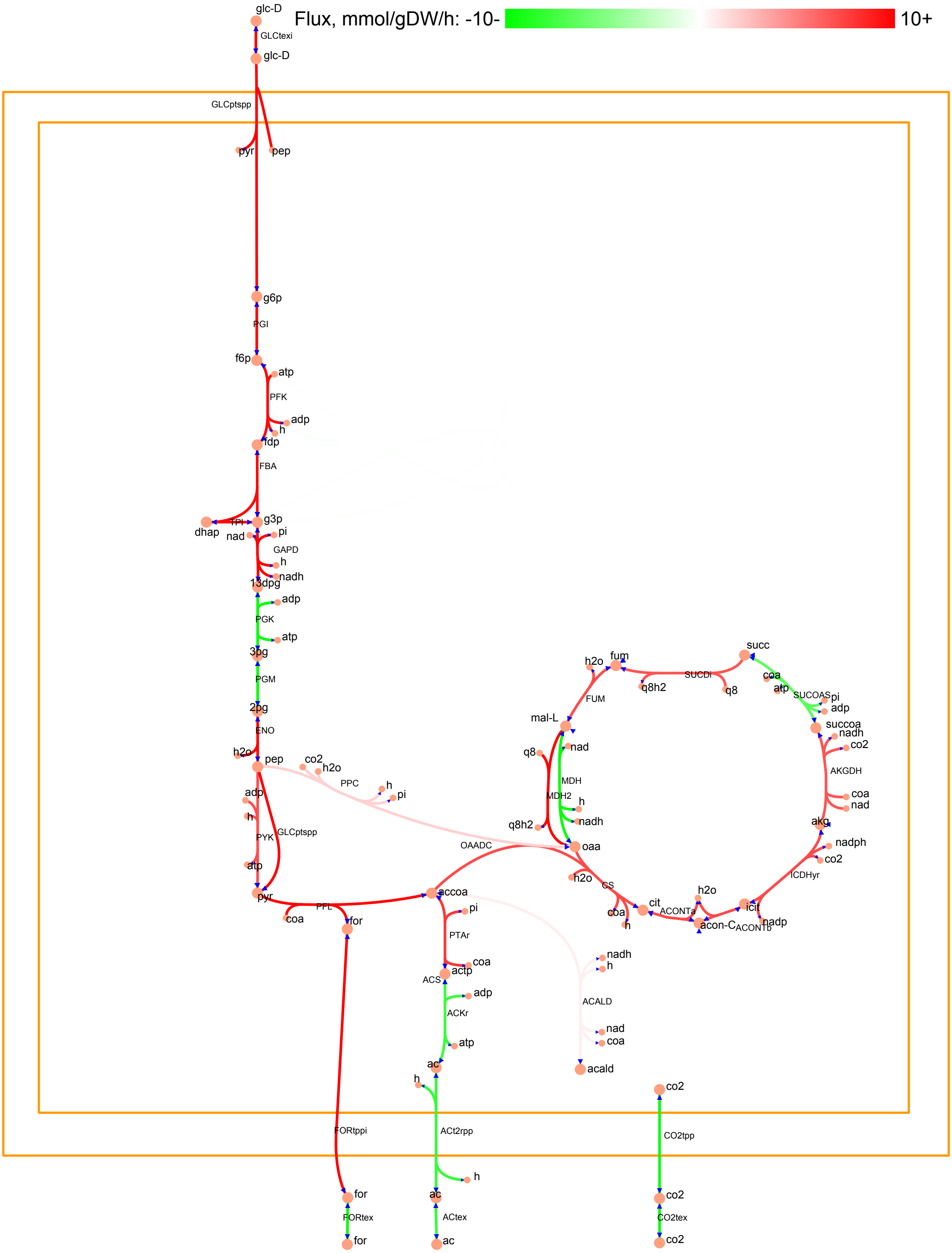

Figure 4.2: MADE FBA core metabolism flux predictions for 3.5 % salt.

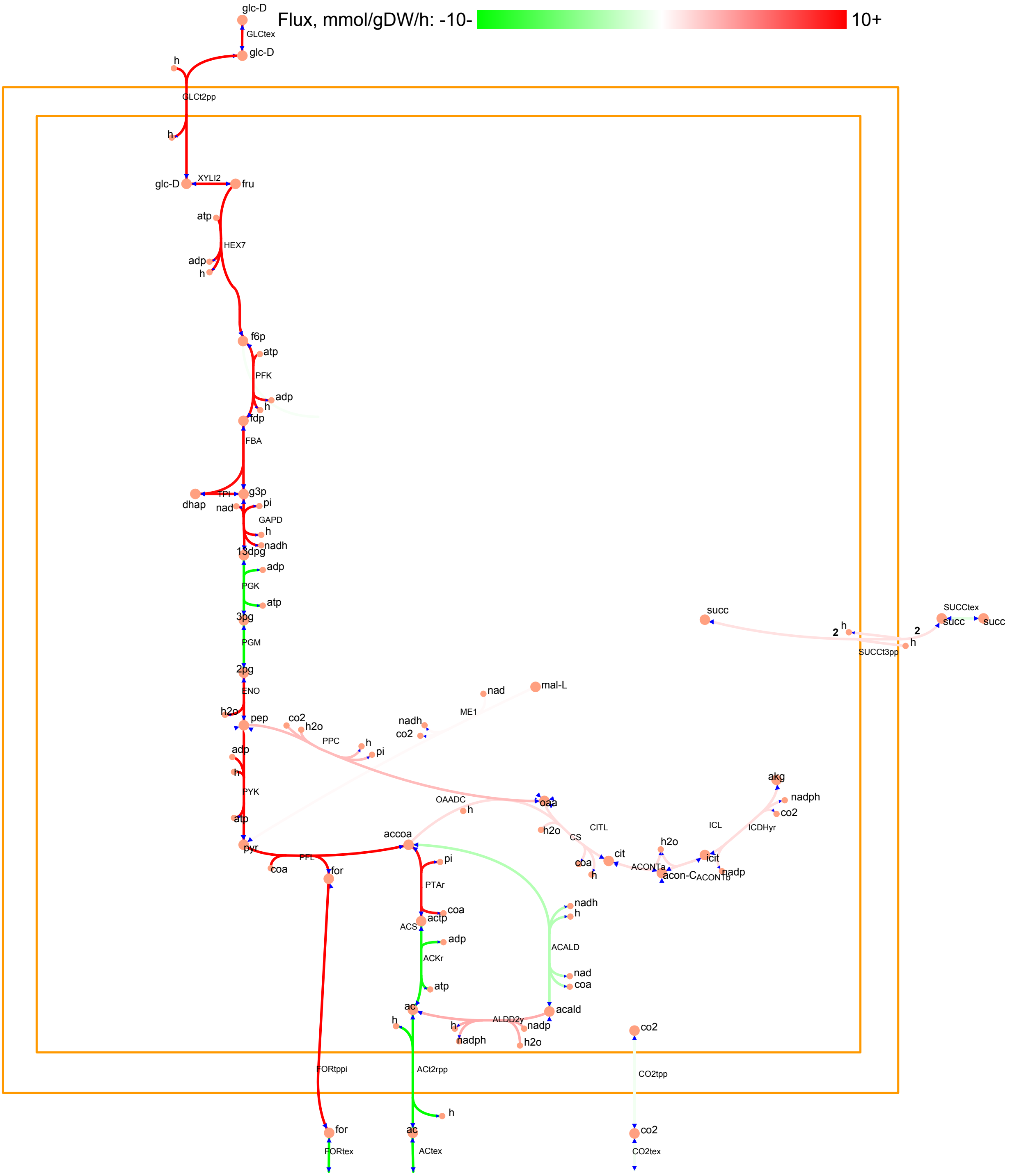

Figure 4.3: MADE FBA core metabolism flux predictions for 4.5 % salt.

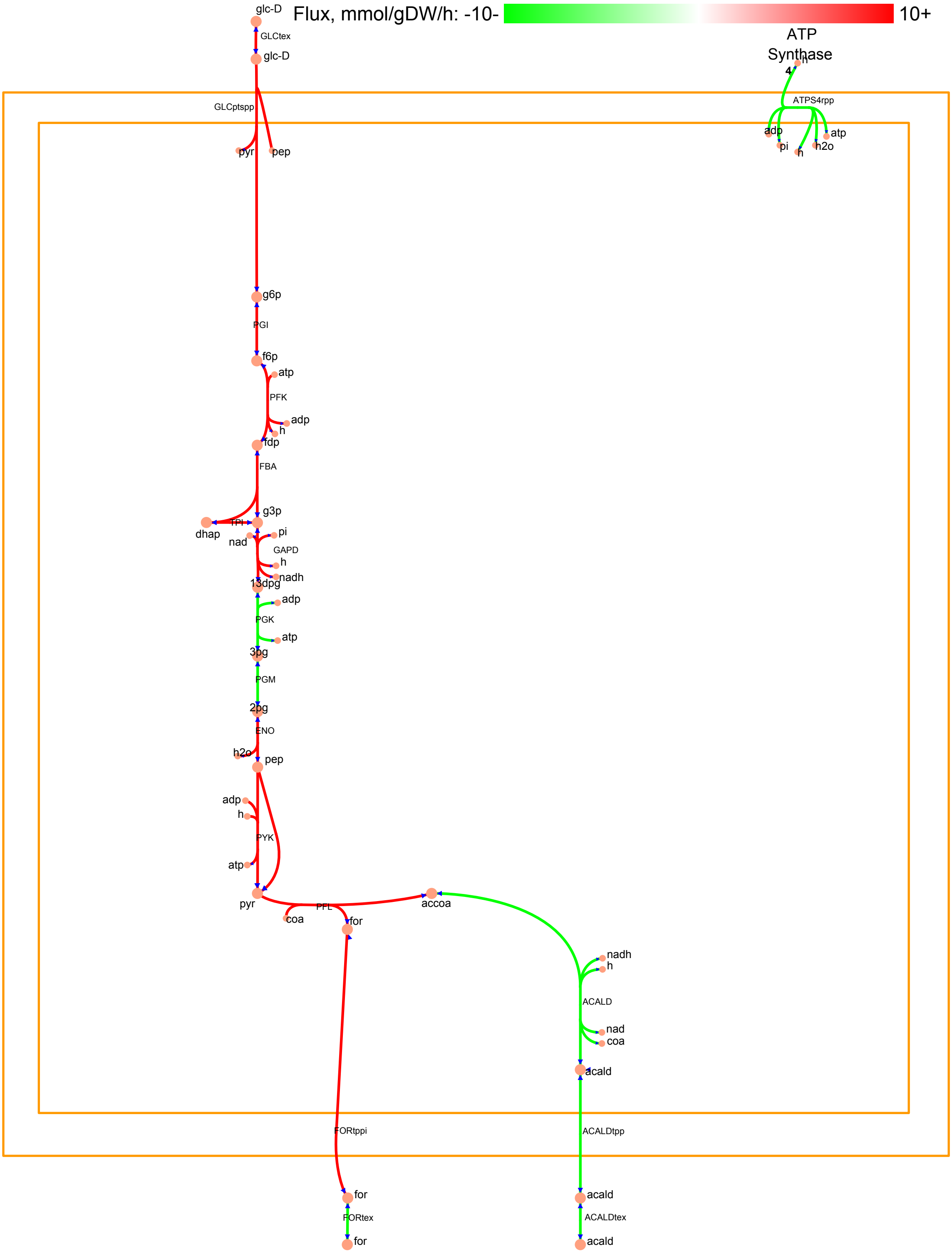

Figure 4.4: MADE FBA core metabolism flux predictions for 5.0 % salt.

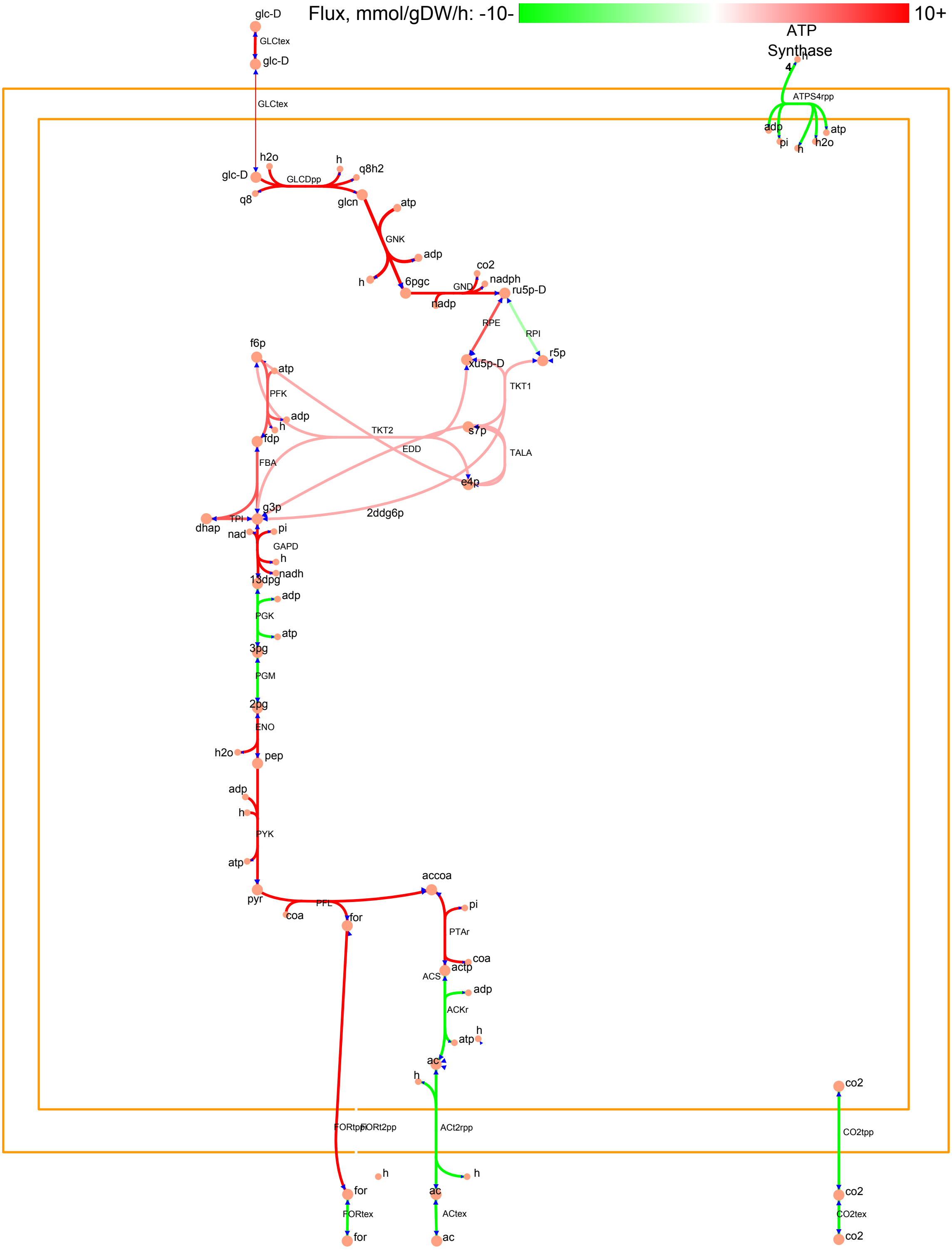

Figure 4.5: MADE FBA core metabolism flux predictions for 5.5 % salt.

At 2.0 % salt, seen in Figure 4.1, MADE predicts that glucose is converted to gluconate in the periplasm of the cell, before its subsequent import into the cytoplasm. In the cytoplasm, gluconate is converted to 6-phosphogluconate (6pgc) by gluconate kinase (GNK). The majority of the gluconate is transformed into glyceraldehyde–3–phosphate (g3p) and phosphoenolpyruvate (pep). Instead of progressing through glycolysis, half of the g3p isomerizes into dihydroxyacetonephosphate (dhap). Following the isomerization, g3p and dhap react to create fructose-bisphosphate, which is converted across two more reactions into glucose–6–phosphate (g6p). G6p completes the cycle, being converted into 6pgc, producing 1 mol NADPH pr mol 6pgc. This bypass of glycolysis allows the cell to produce NADPH instead of the NADH which would be produced during "normal" glycolysis, providing the cell with valuable reducing agents. This is an indication that the cell is experiencing oxidative stress, which has been linked to osmotic stress in a number of studies.[3,39,40,41] Large parts of the Citric Acid Cycle do not carry flux at 2.0 % salt. Those that do, also produce NADPH, as oxaloacetate (oaa) is ultimately converted to alpha-ketoglutarate (akg). The high flux reactions from fumarate to oxaloacetate are a result of a reaction cycle (not shown in the figure) involving an intermediate L-aspartate, creating artificially high flux. The majority of the carbohydrates produced by catabolizing gluconate/glucose actually finds its way into the acetate export pathway, where it is removed from the cell. This is another sign of an oxidative stress state, as fermentation products such as acetate are a hallmark of overflow metabolism.[88]

At 3.5 % salt, Figure 4.2, the NADPH generating cycle which bypasses glycolysis is no longer active. Instead, the Citric Acid Cycle is fully operational, possibly meeting the cells demand for NADPH. The signs of oxidative stress are strenghtened, as the cell begins exporting formate in addition to acetate, both of which are fermentation products.[88]

For cells in 4.5 % salt, Figure 4.3, large parts of the Citric Acid Cycle are dismantled, and the remaining reactions allow flux up to alpha–ketoglutarate (akg). Aside from resulting in the production of NADPH, allowing the Citric Acid Cycle to continue up to akg might be beneficial for the akg itself. Akg is a known precursor for the production of glutamine, a known osmoprotectant.[2] An interesting difference from 3.5 % salt is that acetaldehyde (acald) is produced from acetyl–CoA (accoa), consuming NADH but gaining NADPH. The resulting acetate combines with the acetate produced from accoa, and is exported. Involving acetaldehyde in this way has no effect except for the aforementioned consumption/production of NADH/NADPH, and so it is possible that this is done to compensate for the lowered activity in the NADPH producing reactions of the Citric Acid Cycle. The apparent demand for NADPH, combined with the continued export of fermentation products such as formate and acetate support the claim that oxidative and osmotic stress are inherently linked.[3,39,40,41]

In Figure 4.4, depicting the central metabolic flux distribution of *E. coli* at 5.0 % salt, the Citric Acid Cycle is completely inactive. The shutdown of the Citric Acid Cycle between 4.5 % and 5.0 % salt is in line with observation made in other studies. Metris et. al. reported that there is a shift from aerobic to fermentative metabolism between 4.5 % and 5.0 % salt – a similar response to that observed during oxidative stress.[3] MADE makes a curious prediction here, setting the flux through the ATP synthase reaction to negative. This would mean that ATP is being actively consumed in order to pump protons across the cellular membrane and into the periplasm. If correct, this could mean that cells undergoing severe osmotic stress are able to counteract desiccation by reversing the directionality of the ATP synthase reaction. Studies done on *Campilobacter jejuni* (*C. jejuni*) showed that hyperosmotic stress induces the expression of the ATP synthase gene.[89] Another interesting change at 5.0 % salt is that acetate export has stopped completely, being supplanted by acald export. In light of the predicted reversal of ATP synthase flux, it is difficult to see how shutting down acetate export is beneficial. Notice in Figure 4.3 that the acetate export pathway actually results in the production of ATP. It is reasonable to assume that the reverse ATP synthase activity would benefit from having as many ATP producing reactions active as possible, making this an odd prediction.

At 5.5 % salt, Figure 4.5, the ATP synthase reversal is maintained. Glucose enters through the Entner Doudoroff (ED) pathway, consuming ATP but producing NADPH along its catabolic path. As was observed at 2.0 % salt, the products from the ED pathway eventually end up as g3p and fructose, albeit through different reactions. From here, flux proceeds in a completely linear fashion down to pyruvate, which branches off into either acetyl–CoA or formate. Formate is exported from the cell, along with acetate, which has had its export reaction restored, resuming its likely valuable production of ATP for the cell. It is important to note that MADE predicted no growth for *E. coli* in both 5.0 % and 5.5 % salt, which could be a result of the reverse ATP synthase reaction. If true, these predictions could indicate the existance of a stationary phase phenotype during osmotic stress where almost all energy is devoted towards osmoadaptation.

Looking at the whole picture, there is a trend of increasing where the export of fermentation products increases with increasing salt concentrations. At some point, between 4.5 % and 5.0 % salt, there is a fundamental shift in metabolism, where the Citric Acid Cycle shuts down and almost ATP generated during glycolysis is directed toward pumping protons from the cytoplasm into the periplasm. Interestingly, this phenomenon has been reported previously in cancer cells, where the release of cytochrome-*c* induced a similar response in mitochondrial ATP synthases.[90] Additonaly, Perroud and Rudulier reported that transport of glycine betaine is driven by the electrochemical proton gradient in *E. coli.*[91] Glycine betaine is an effective osmoprotectant in *E. coli*, and *E. coli* cells growing in its presence during osmotic stress have significantly improved growth rates.[3]

4.2 Combination of OMICS data and MUTE method

The MUTE method relies on gene expression data for which there is overlapping data on gene expression levels, translation efficiencies, protein molecular weights and enzyme turnover rates. Some of these parameters, such as enzyme turnover rates and translation efficiencies, are difficult to measure, and as a result these kinds of data only exist for a small subset of all genes in organisms such as *E. coli.*

During MUTE's development, the set of genes for which all of these parameters were known counted only 319 for *E. coli*, or approximately 7% of its protein coding genes. Additionaly, none of the 319 genes were represented in the set of 100 genes who's expression changed the most between salt concentrations. This is currently a limitation for the MUTE method, which limits its range of predictions but, as more data on enzyme turnover rates and mRNA translation efficiencies is gathered, the situation will hopefully improve.

Removing the requirement for overlapping data from gene expression, enzyme turnover rates, translation efficiency and molecular weights increased the number of included genes significantly, from 319 to 905. The loss of accuracy in enzyme turnover rates and translation effiencies were seen as a compromise in order to increase the number of included genes.

To differentiate between the MUTE models generated from the two separate data sets, they will henceforth be referred to as "MUTE models with strict data filtering" for the data set which included 319 genes, and "MUTE models without strict data filtering" for the data set with 905 genes.

In order to visualize the resulting flux bounds placed on the models by the MUTE method, the flux maps in Figures A.0.1 through A.0.5 and B.0.1 to B.0.5 in the appendix were made.

4.3 Comparison of predicted protein concentrations and copy numbers with empirical data

The mean protein copy number and concentration for the five models with and without strict data filtering generated by MUTE is shown in Table 4.1 and 4.2, respectively.

Table 4.1: Mean protein copy numbers and concentrations for the five osmotic stress models generated by MUTE with strict data filtering.

Salt concentration ,%	protein copy number, #	protein concentration, $\frac{\text{mmol}}{\text{gDW}}$
2.0	$2 \cdot 10^3$	$1.3 \cdot 10^{-5}$
3.5	$2 \cdot 10^3$	$1.2 \cdot 10^{-5}$
4.5	$2 \cdot 10^3$	$8.8 \cdot 10^{-6}$
5.0	$2 \cdot 10^3$	$1.3 \cdot 10^{-5}$
5.5	$2 \cdot 10^3$	$1.3 \cdot 10^{-5}$

Table 4.2: Mean protein copy numbers and concentrations for the five osmotic stress models generated by MUTE without strict data filtering.

Salt concentration, %	protein copy number, #	protein concentration, $\frac{\text{mmol}}{\text{gDW}}$
2.0	$1 \cdot 10^3$	$6.2 \cdot 10^{-6}$
3.5	$1 \cdot 10^3$	$6.0 \cdot 10^{-6}$
4.5	$1 \cdot 10^3$	$5.7 \cdot 10^{-6}$
5.0	$1 \cdot 10^3$	$6.2 \cdot 10^{-6}$
5.5	$1 \cdot 10^3$	$6.1 \cdot 10^{-6}$

Tables 4.1 and 4.2 show that the average copy number for proteins predicted by MUTE have an order of magnitude of 10^3, which conforms to experimentally measured average protein abundance numbers.[10] Notice that the protein copy numbers in Table 4.2 are lower than in Table 4.1, resulting

from the difference in the number of included genes. Predicting realistic protein copy numbers is an important step towards generating realistic simulations of cell metabolism, but in order to make meaningful predictions it is important that the predicted protein abundance distributions match experimentally measured ones. The protein copy number distributions is shown in Figures 4.6 and 4.7 for MUTE models with and without strict data filtering, respectively.

The protein copy number distributions in Figures 4.6 and 4.7 show an enrichment of proteins with abundances around $8 \cdot 10^3$ copies/cell. These enrichments are an artifact of a constraint on all protein concentrations, which limits the maximum protein concentration to 20 $\frac{\mu\,\text{mol}}{\text{gDW}}$. Without this constraint, it is reasonable to assume that the distribution would lie closer to an exponential distribution. This constraint was imposed as a way of evening out the protein distribution, as in its absence genes with very high expression quickly depleted the cell's enzyme mass budget, leaving most enzymes with negligible concentrations. The choice of 8000 copies pr cell was chosen as it is approximately twice the size of the average protein concentration reported in the litterature, allowing some room for "extreme expression".[10]

Previous studies have reported that protein abundance distributions can be described by a gamma distribution.[92] Looking at Figures 4.6 and 4.7, the protein abundance distributions resemble a combination of an exponential and a flat distribution. Proteins with up to 3000 copies pr cell follow an exponential distribution while proteins above 3000 copies pr cell appear to follow a flat distribution. Keeping in mind that the exponential distribution is a particular case of the gamma distribution, the predicted protein distributions fit empirical data partially, although not completely.

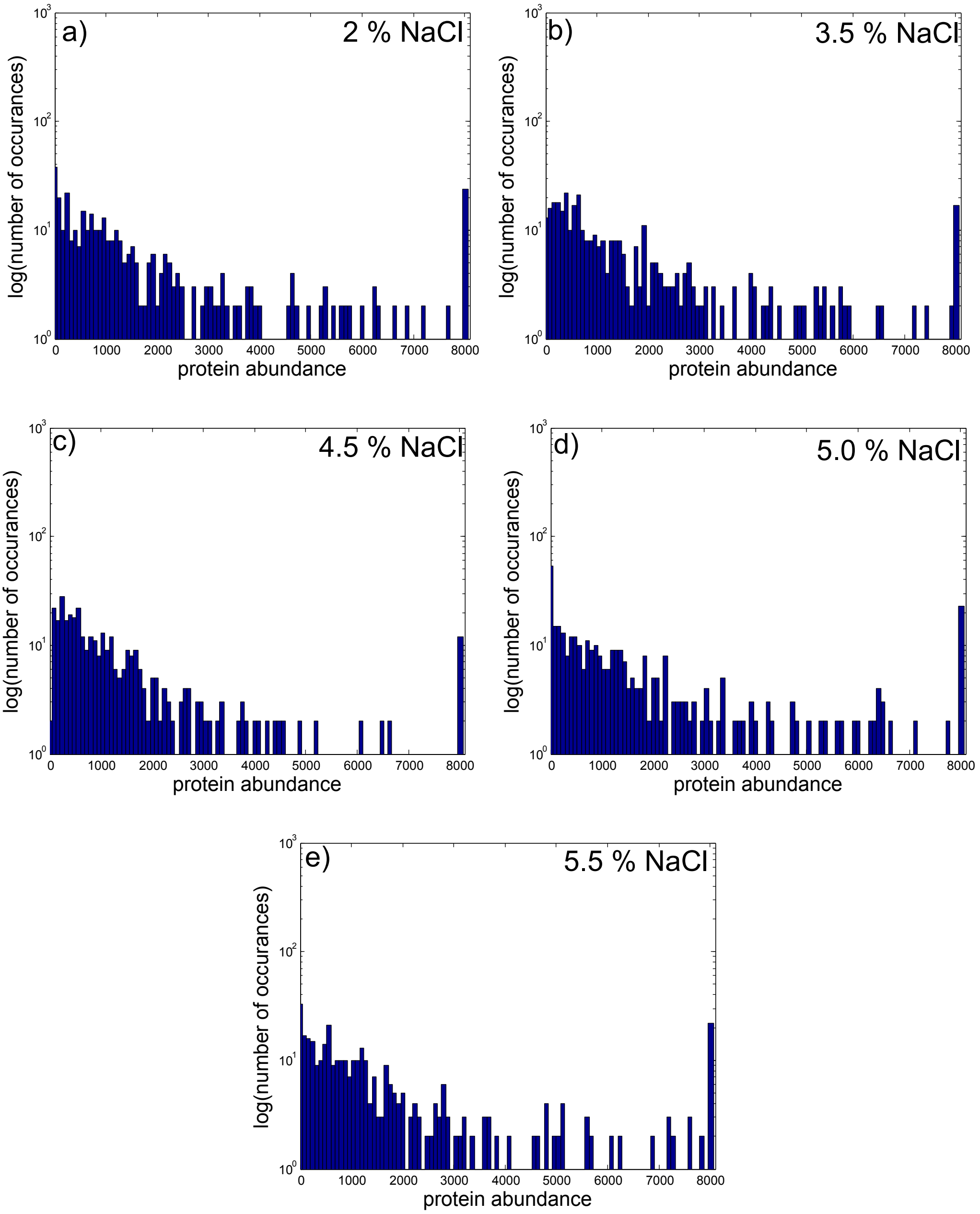

Figure 4.6: Protein copy number distributions for all MUTE models with strict data filtering.

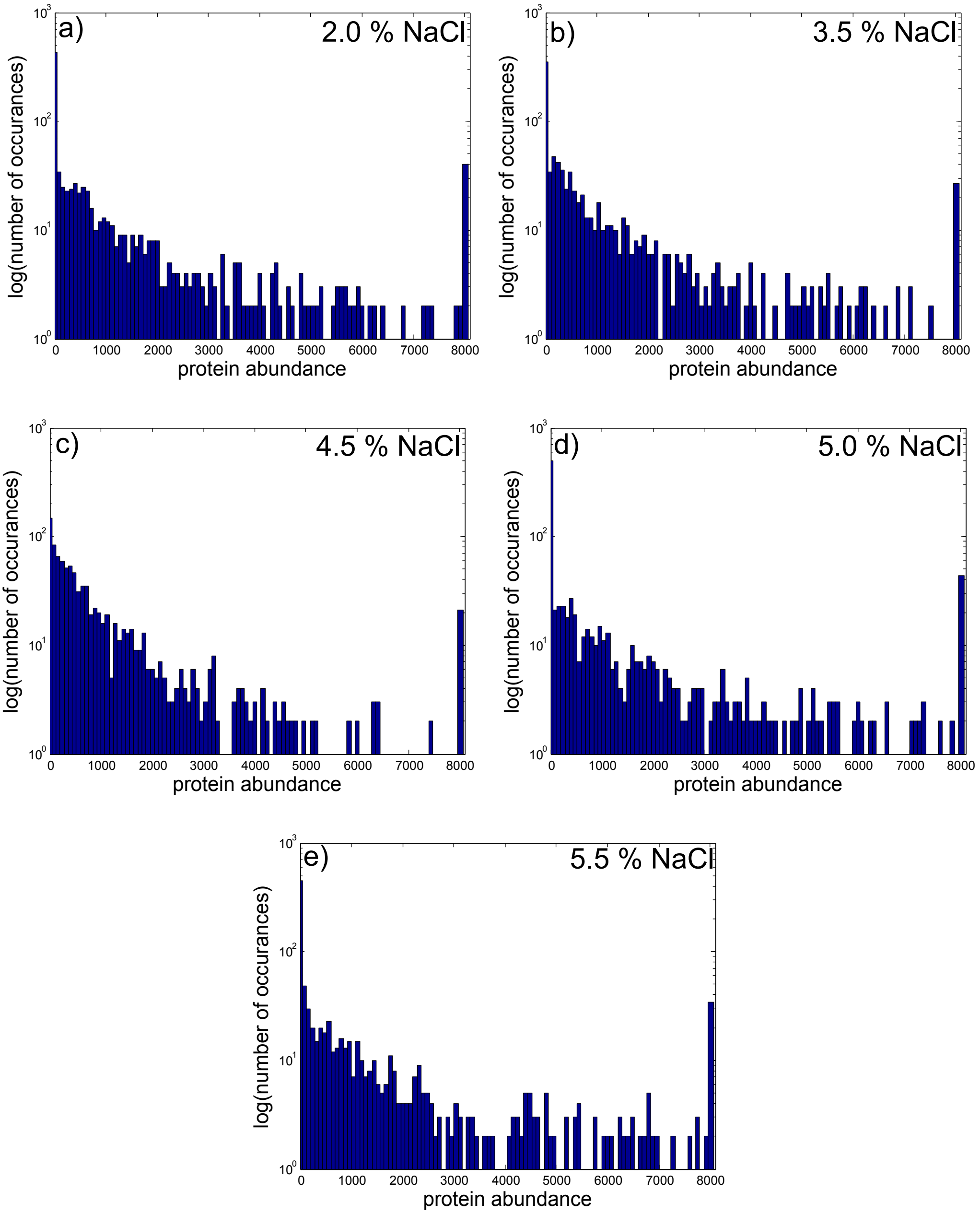

Figure 4.7: Protein copy number distributions for all MUTE models without strict data filtering.

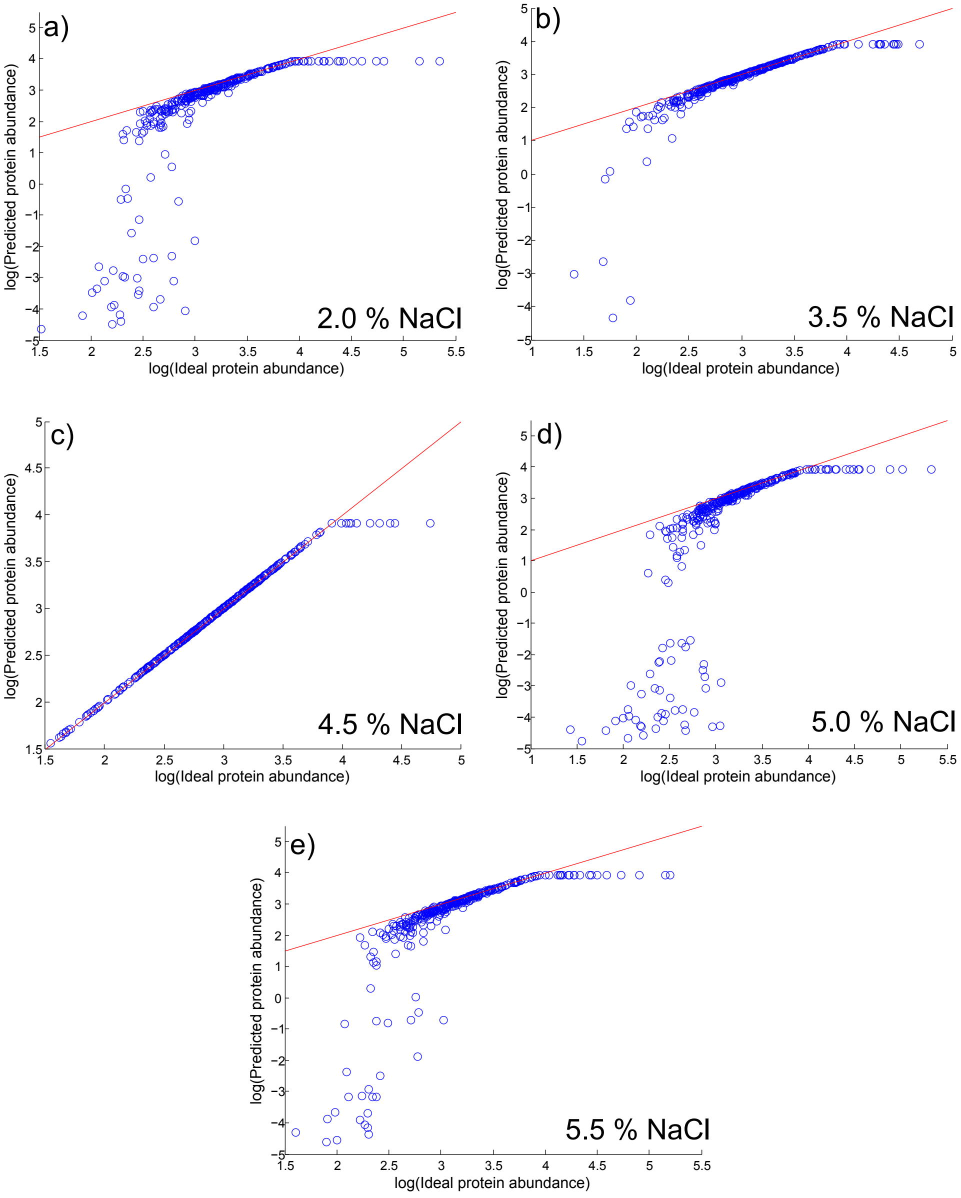

Figure 4.8: Protein abundances for MUTE models with strict data filtering. Without the total enzyme mass constraint of MUTE, all proteins would align with the red line. Axes are logarithmic.

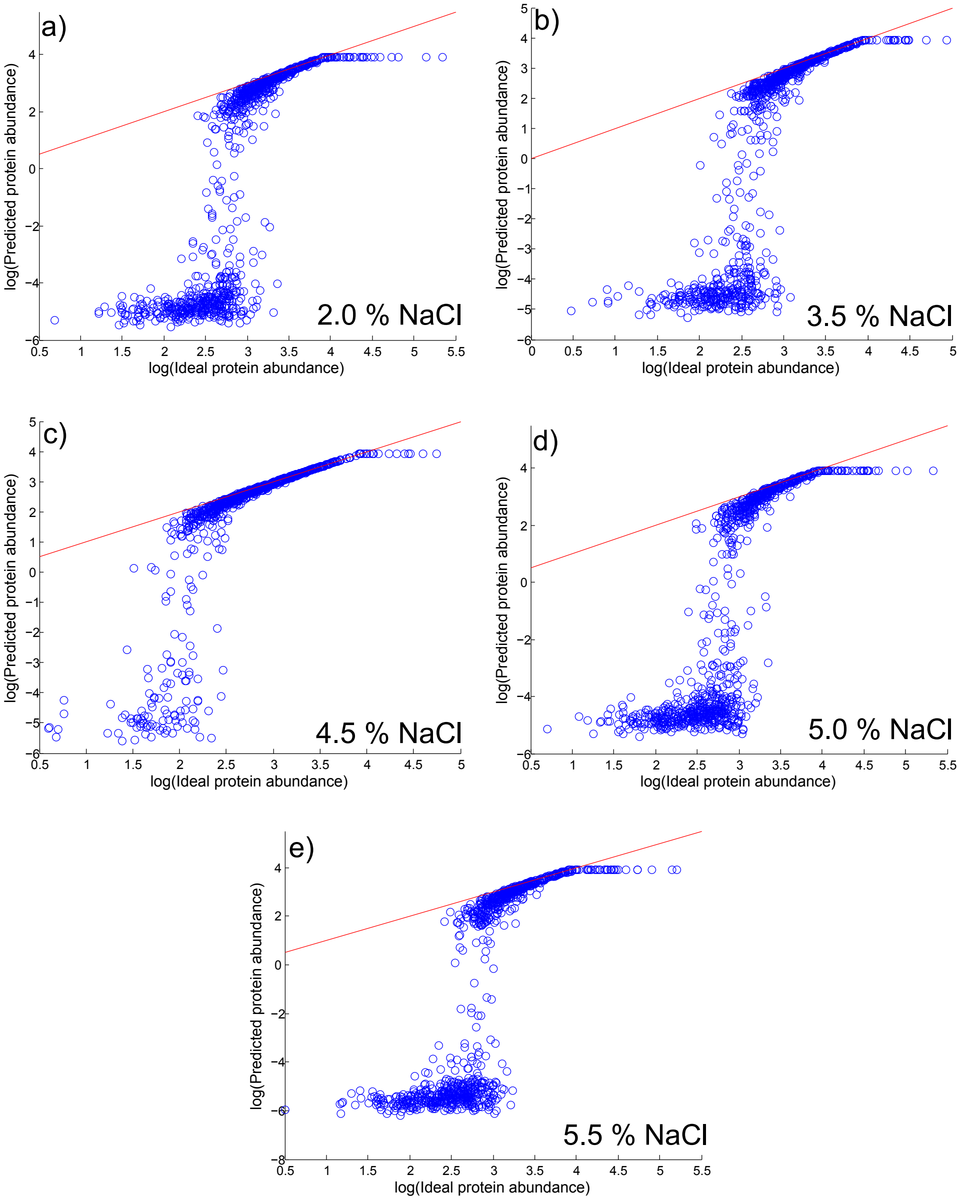

Figure 4.9: Protein abundances for MUTE models without strict data filtering. Without the total enzyme mass constraint of MUTE, all proteins would align with the red line. Axes are logarithmic.

Figures 4.8 and 4.9 visualize the effects of the total enzyme mass constraint placed on MUTE. If no total enzyme mass constraint was present, all protein abundances would align to the red line displayed in the figures. With the constraint active, the optimization step of MUTE distributes mass to the enzymes, prioritizing those who have higher gene expression and translation efficiency values. In Figure 4.8, a large fraction of the proteins are able to achieve their desired concentrations, while in Figure 4.9 this fraction is considerably lower. This stems from the difference in the amount genes included in the input data to MUTE, where the MUTE models without strict data filtering contain data on 3 times as many genes.

Notice in both data filtering regimes, that at 4.5 % salt the enzyme mass demand from gene expression appears to be much lower than in the other cases. This can be seen in Figure 4.8c, where the data points almost perfectly line up to the red line, and in Figure 4.9c where the number of proteins without their "desired" abundance is much lower. This will have the effect of softening the effect of the total enzyme mass constraint used by MUTE during the protein concentration optimization process, and in extreme cases a low enough total gene expression could make the total enzyme mass constraint obsolete. In these extreme situations, even genes with very low expression are "translated" into proteins.

4.4 Unbiased random sampling of MUTE solution spaces

Sampling the solution space of the models produced by MUTE can help identify complex and subtle differences between the predicted metabolisms, but the analysis of the solution spaces only give us measures of possibilities and potentials. In order to test the performance of the MUTE models when "growing" with a defined objective or phenotype, the MUTE models had their objective function flux constrained to the optimum flux predicted by Flux Balance Analysis. The models were then sampled, and the average of the sample points were taken for all reactions, resulting in an average flux vector.

The resulting flux predictions for the core metabolism of *E. coli* is visualized in Figures 4.10 to 4.14 and Figures 4.15 to 4.19 for MUTE models with and without strict data filtering, respectively.

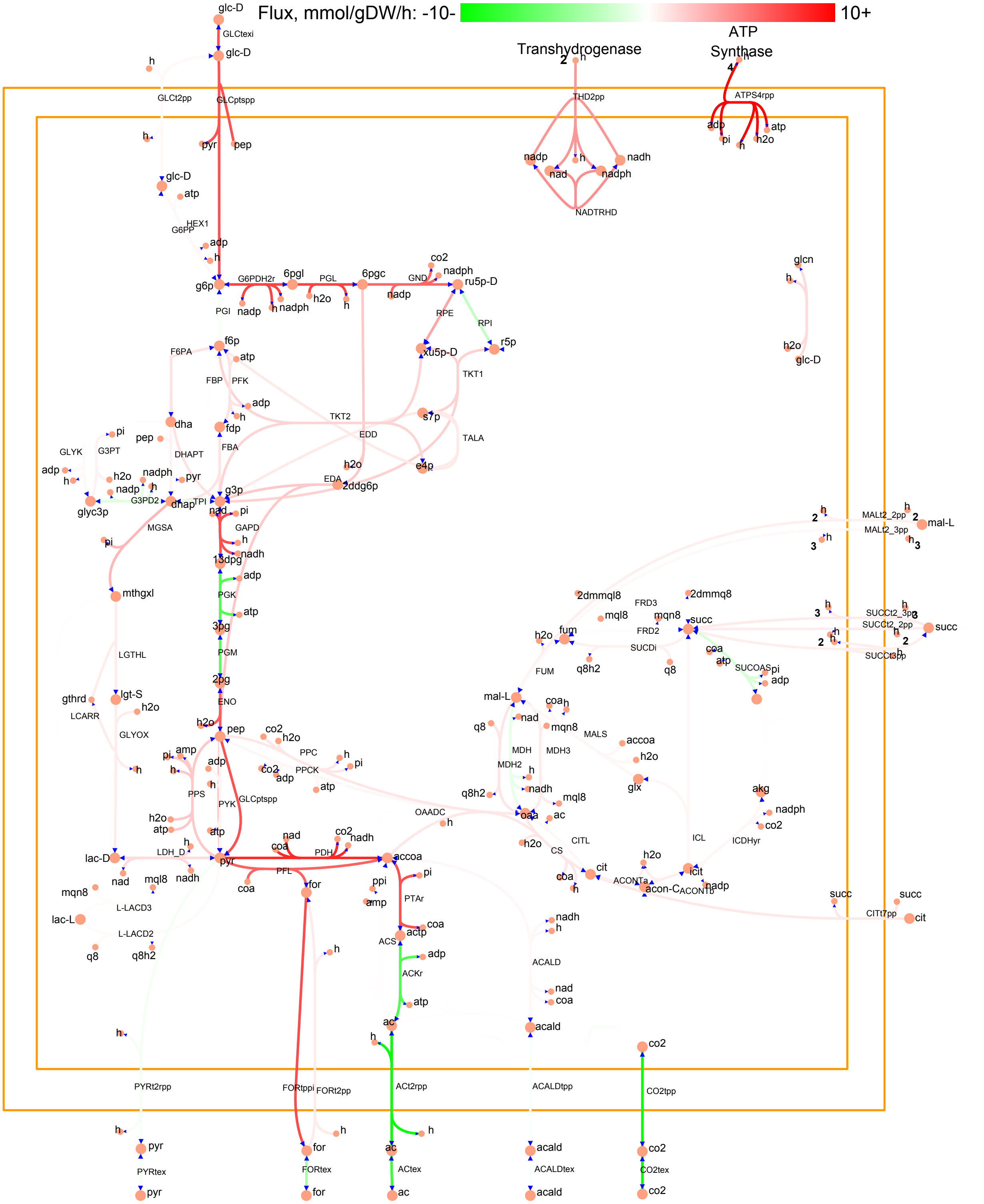

Figure 4.10: MUTE average core metabolism flux predictions for 2.0% salt. Made with strict data filtering.

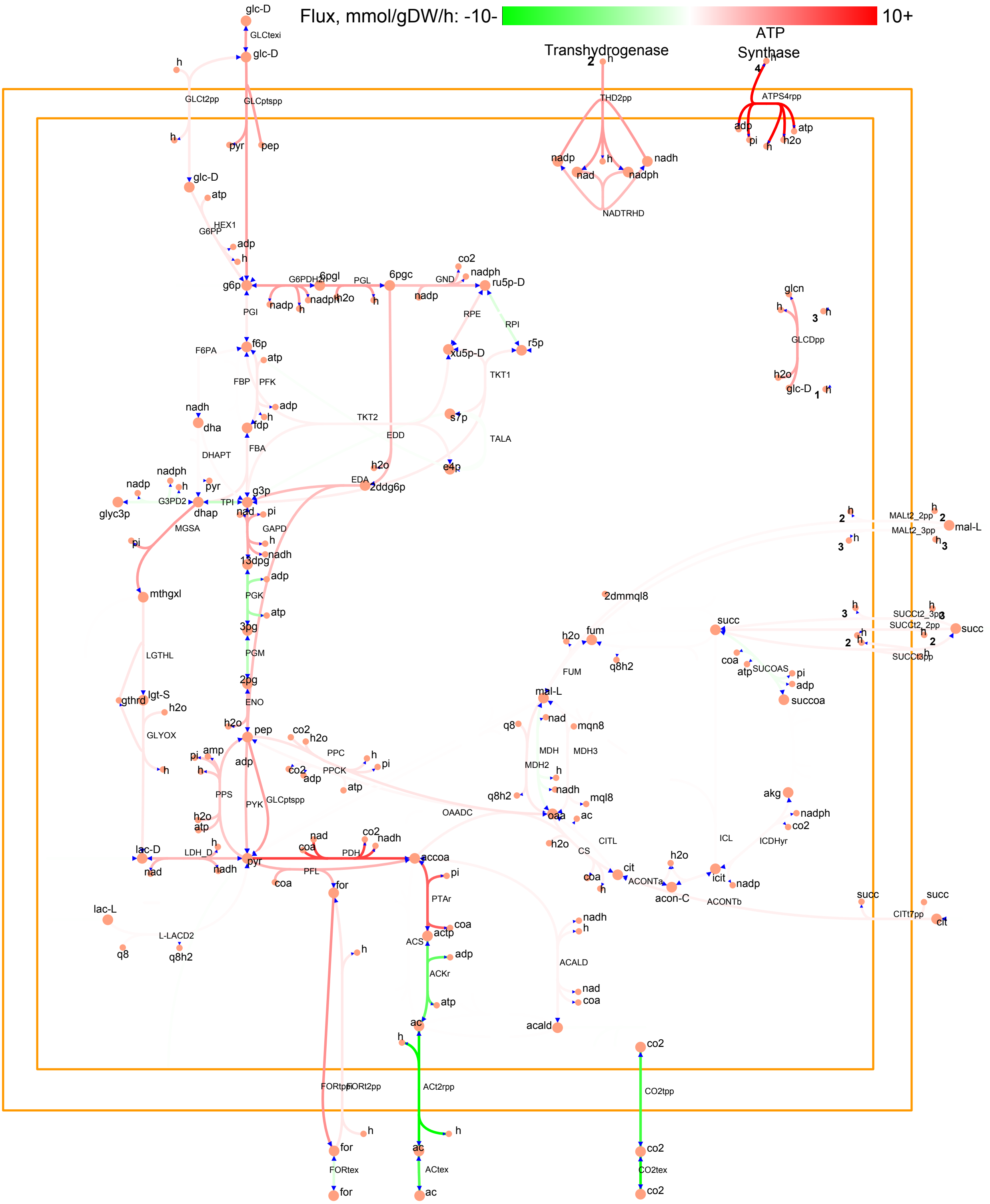

Figure 4.11: MUTE average core metabolism flux predictions for 3.5% salt. Made with strict data filtering.

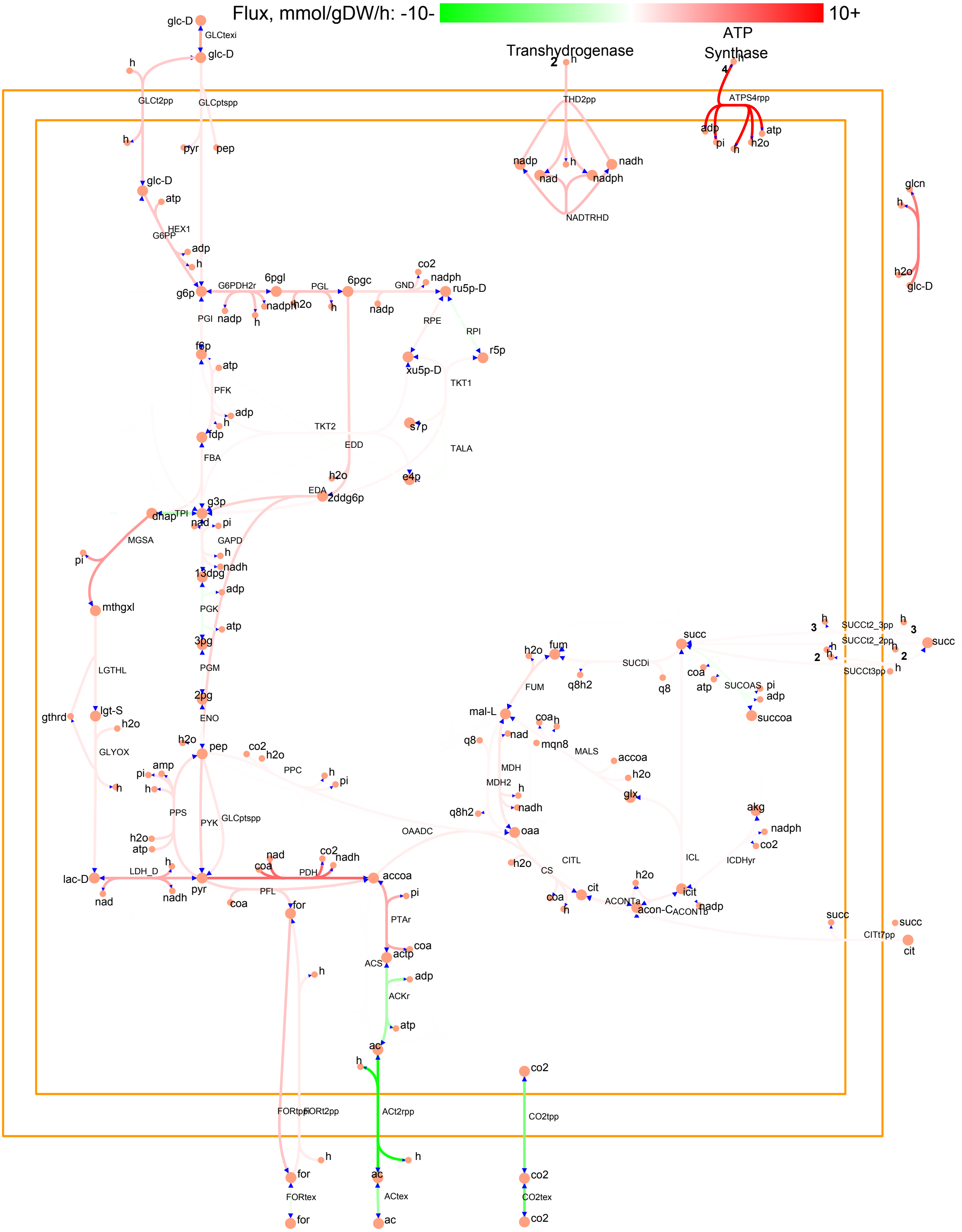

Figure 4.12: MUTE average core metabolism flux predictions for 4.5% salt. Made with strict data filtering.

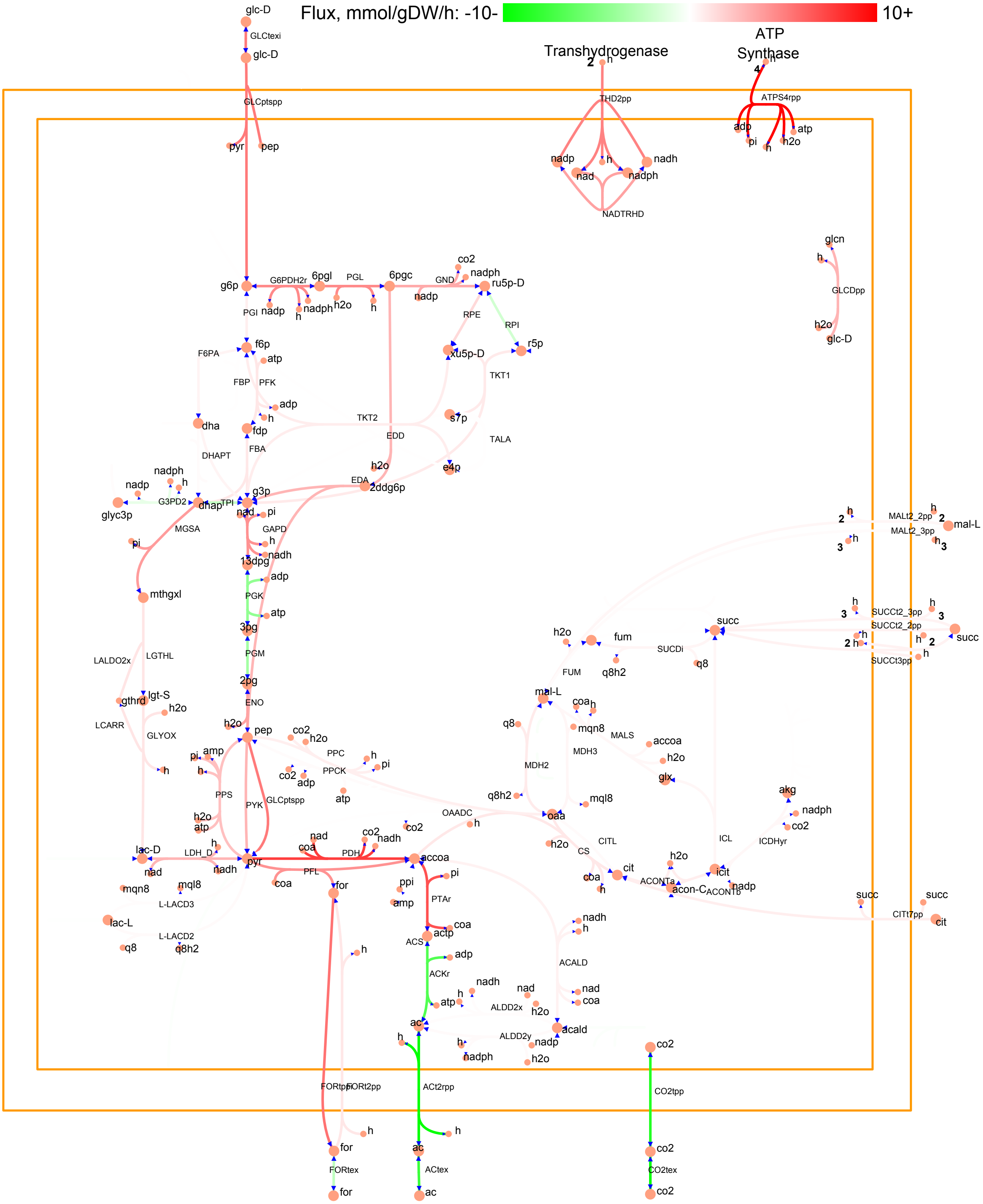

Figure 4.13: MUTE average core metabolism flux predictions for 5.0% salt. Made with strict data filtering.

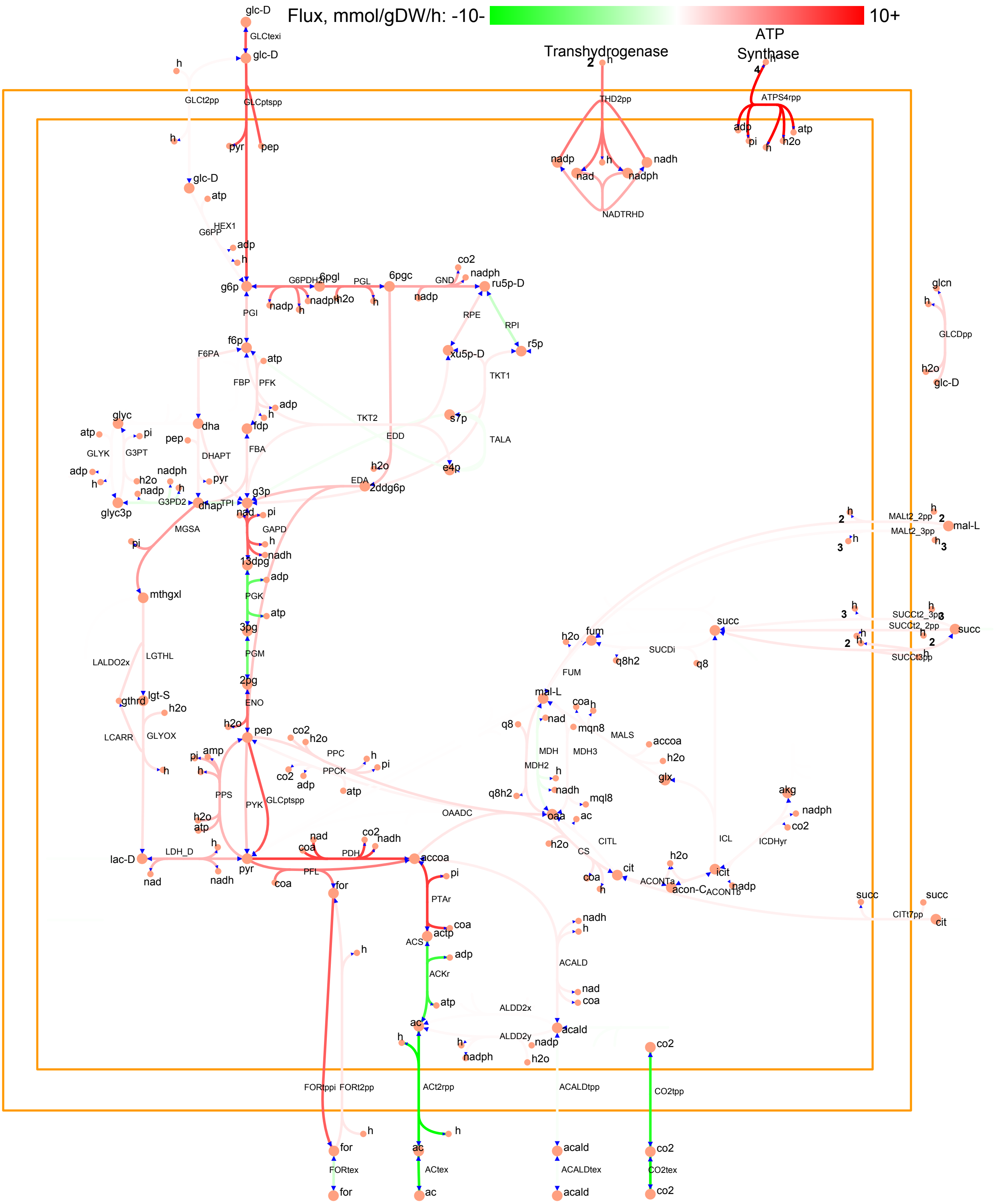

Figure 4.14: MUTE average core metabolism flux predictions for 5.5% salt. Made with strict data filtering.

Beginning with the MUTE models generated with strict data filtering in Figures 4.10 to 4.14, there are big differences when compared to MADE. Much larger parts of the metabolic network is maintained across all models, differing mostly in the strenghts of the fluxes.

In Figure 4.10, glucose enters the cell through the ED pathway, producing 2 NADPH for each glucose consumed on its way towards ribulose–5–phosphate (ru5p–D). From the ED pathway, the carbohydrates flow towards fructose–6–phosphate (f6p) and glyceraldehyde–3–phosphate (g3p). Notice the relatively large flux through the glyoxal bypass, resulting in large amounts of methylglyoxylate (mthgxl). The carbon resources for producing mthglx come from f6p and g3p, both of which are converted to dihydroxyacetone–phosphate (dhap). Mthgxl, which is highly toxic to cells is rapidly converted into D–lactate, through several reactions, some of which are not shown in Figure 4.10. The function methylglyoxal bypass, in light of its potential toxicity is not well understood. Studies in *Saccharomyces cerevisiae* (*S. cerevisiae*) have shown that expression of genes involved in the glyoxal bypass are induced during osmotic stress, and these predictions might indicate that the same response exists in *E. coli.*[93] Methylglyoxylate is rapidly converted into D–lactate through several reactions, some of which are not visible in Figure 4.10. It is converted into pyruvate, supplementing the pyruvate produced through glycolysis. As was the case for the MADE predictions, most of the pyruvate is converted into fermentation products such as formate, acetate and acetaldehyde, which proceed to be exported from the cell. Some flux is diverted from the export of fermentation products and into the Citric Acid Cycle where isocitrate is shuttled through the glyoxlate shunt, completing the cycle. The major export of fermentation products is unexpected, as at 2.0 % salt the cell should not be experiencing major osmotic stress, and by extension oxidative stress.

At 3.5 % salt, shown in Figure 4.11 there are few changes to the connectivity of the metabolic network compared to 2.0 % salt, but there are changes in the flux distribution. Most of the glucose entering the cytoplasm still goes through the ED pathway, but a small amount of glucose–6–phosphate (g6p) is isomerized into f6p, which is converted into g3p and dhap. The flux entering the ED pathway is mostly converted into g3p and pyruvate (pyr), the downstream steps of which are very similar to that of 2.0 % salt. The same diversion of dhap into the glyoxal bypass is observed, eventually ending up as D-lactate. The distribution of fluxes for 3.5 % salt would lead to a decrease in the production of NADPH, contradicting the trend observed from the MADE predictions, where NADPH production was maintained at the cost of ATP. The low activity through the Citric Acid Cycle, as well as the production and export of fermentation products is maintained, further strengthening the hypothesis that osmotic stress induces excretion of fermentation products.

As the salt concentration hits 4.5 % in Figure 4.12 it is difficult to see any changes to the metabolism compared to Figure 4.11. There is less import of glucose, resulting lower flux through the whole network. This seemingly eliminates some reactions which had very low flux at 3.5 % salt, but the flux follows the same pattern: catabolism of glucose in the ED and Embden–Meyerhof–Parnas (EMP) pathways, followed by heavy formate and acetate export, and an almost inactive Citric Acid Cycle. At 4.5 % the cell should be experiencing significant osmotic stress, and it was expected that a shift in the central metabolism at would reflect that. It is possible that the 319 genes who's expression make up the basis of the MUTE predictions simply do not affect the central metabolism in any significant way, and that more data is needed in order to make meaningfull predictions. Looking at Figures 4.13 and 4.14 goes a long way towards confirming this suspicion, as the same pattern emerges. The glucose import flux varies, influending the downstream flux strengths in all reactions, but the identity and connectedness of the active reactions fail to change in any significant way.

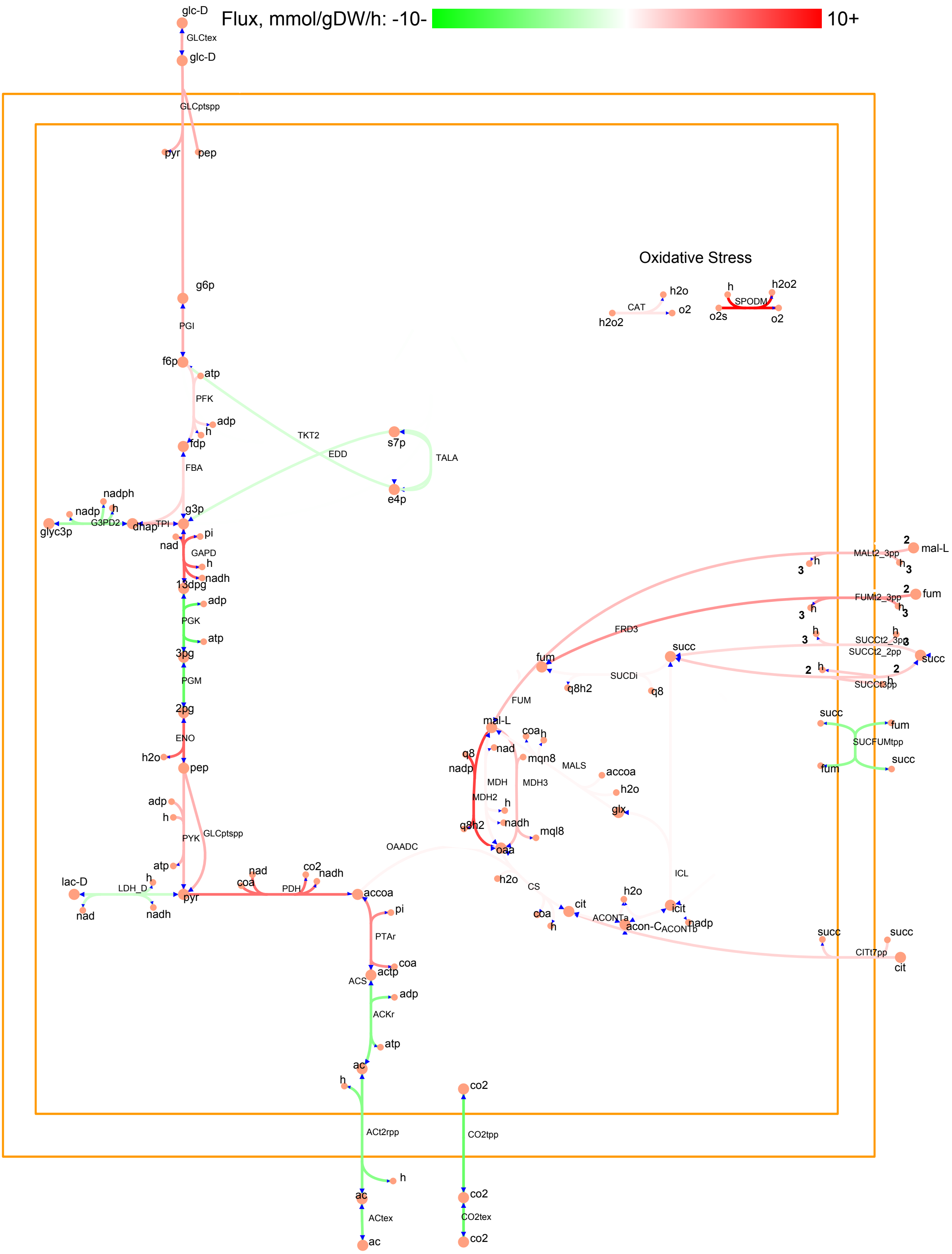

Figure 4.15: MUTE average core metabolism flux predictions for 2.0% salt. Without strict data filtering.

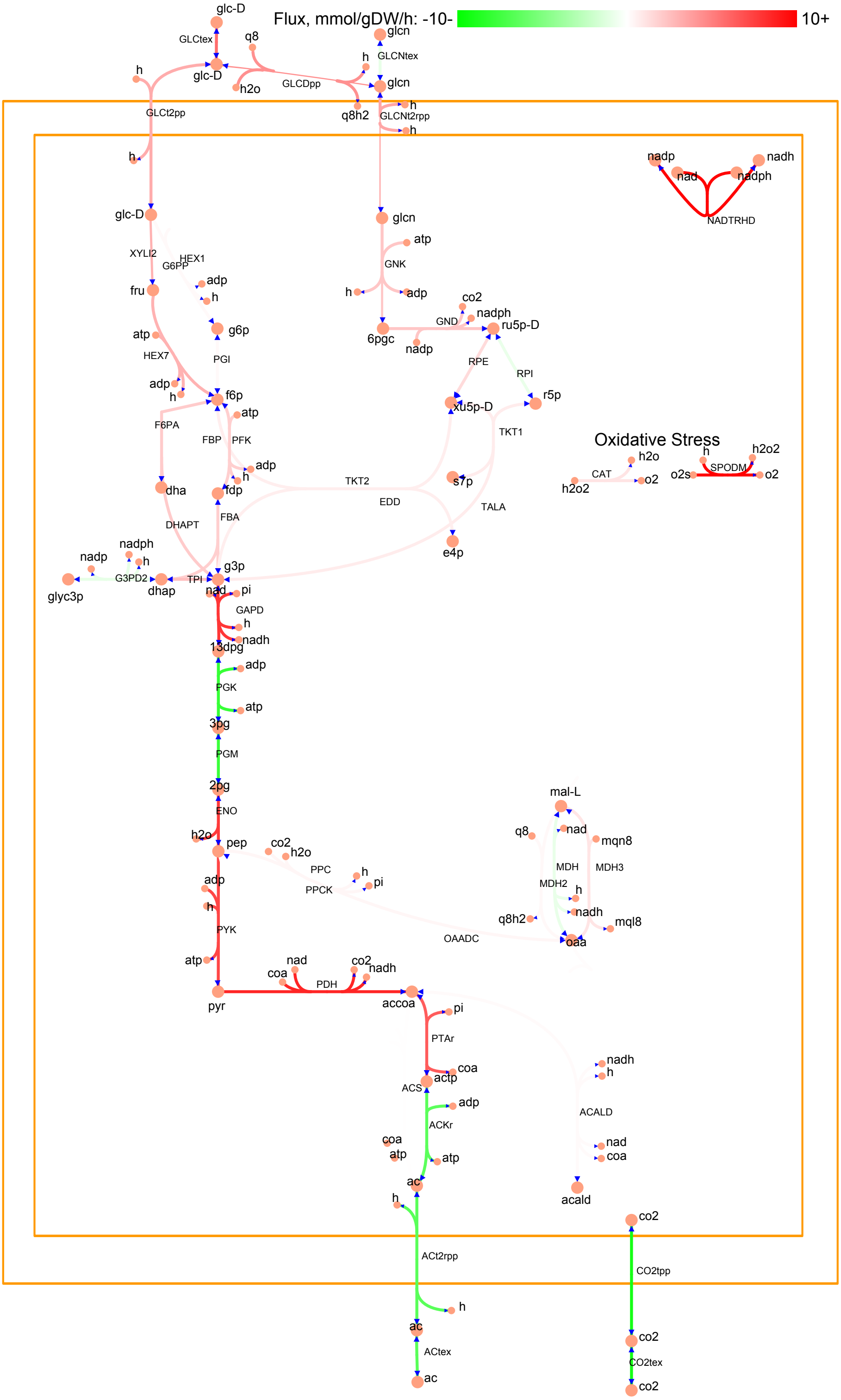

Figure 4.16: MUTE average core metabolism flux predictions for 3.5% salt. Without strict data filtering.

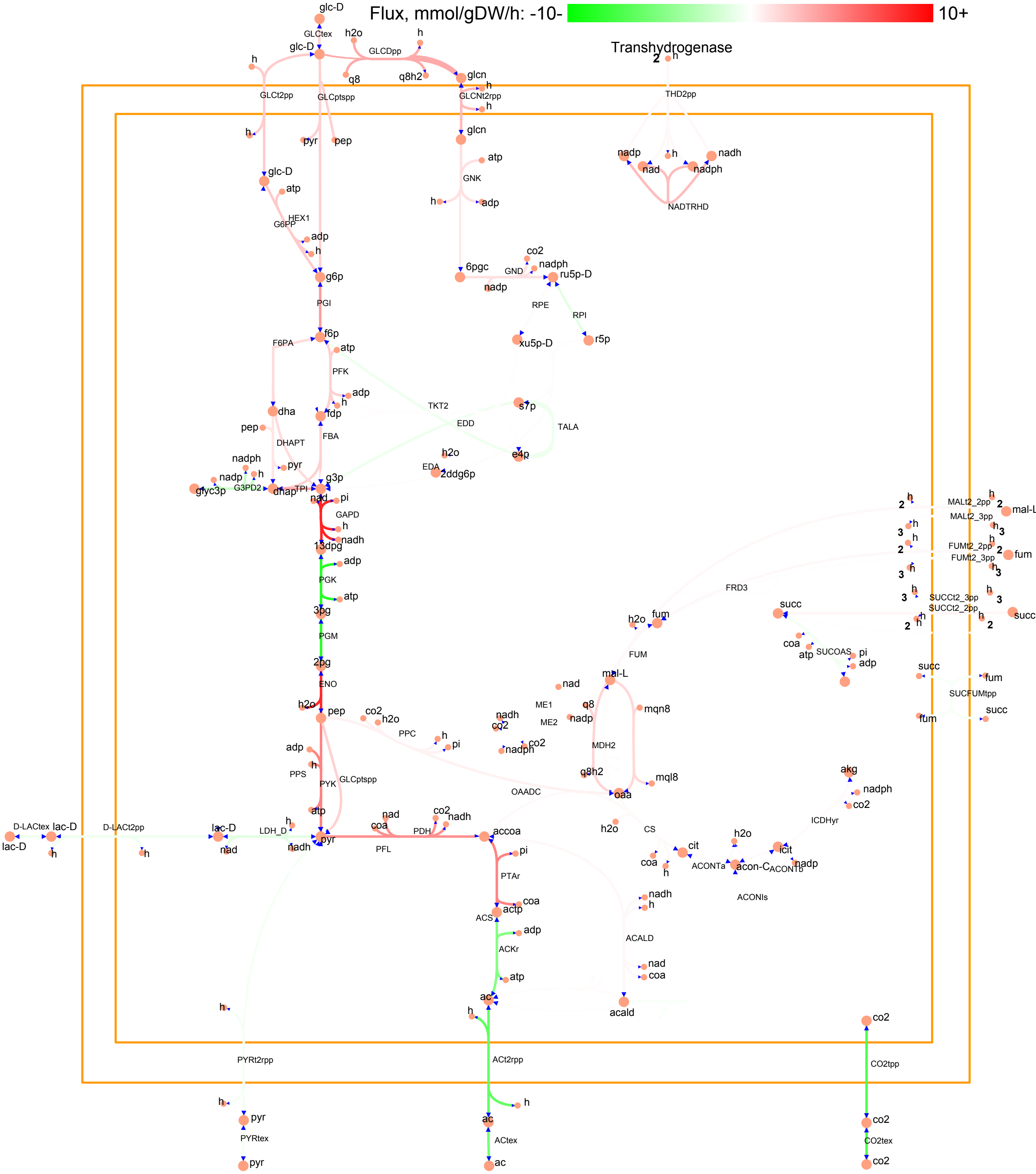

Figure 4.17: MUTE average core metabolism flux predictions for 4.5% salt. Without strict data filtering.

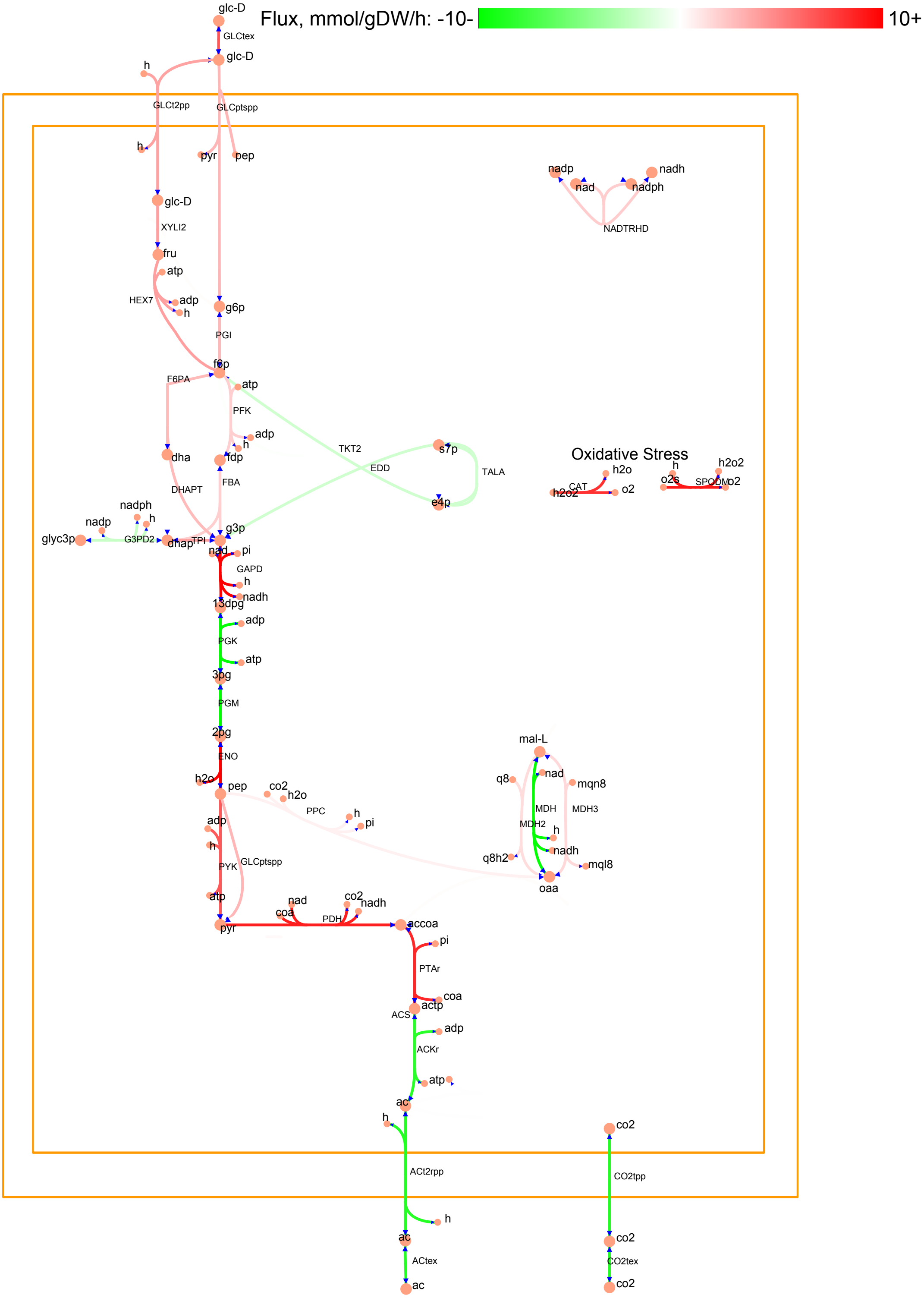

Figure 4.18: MUTE average core metabolism flux predictions for 5.0% salt. Without strict data filtering.

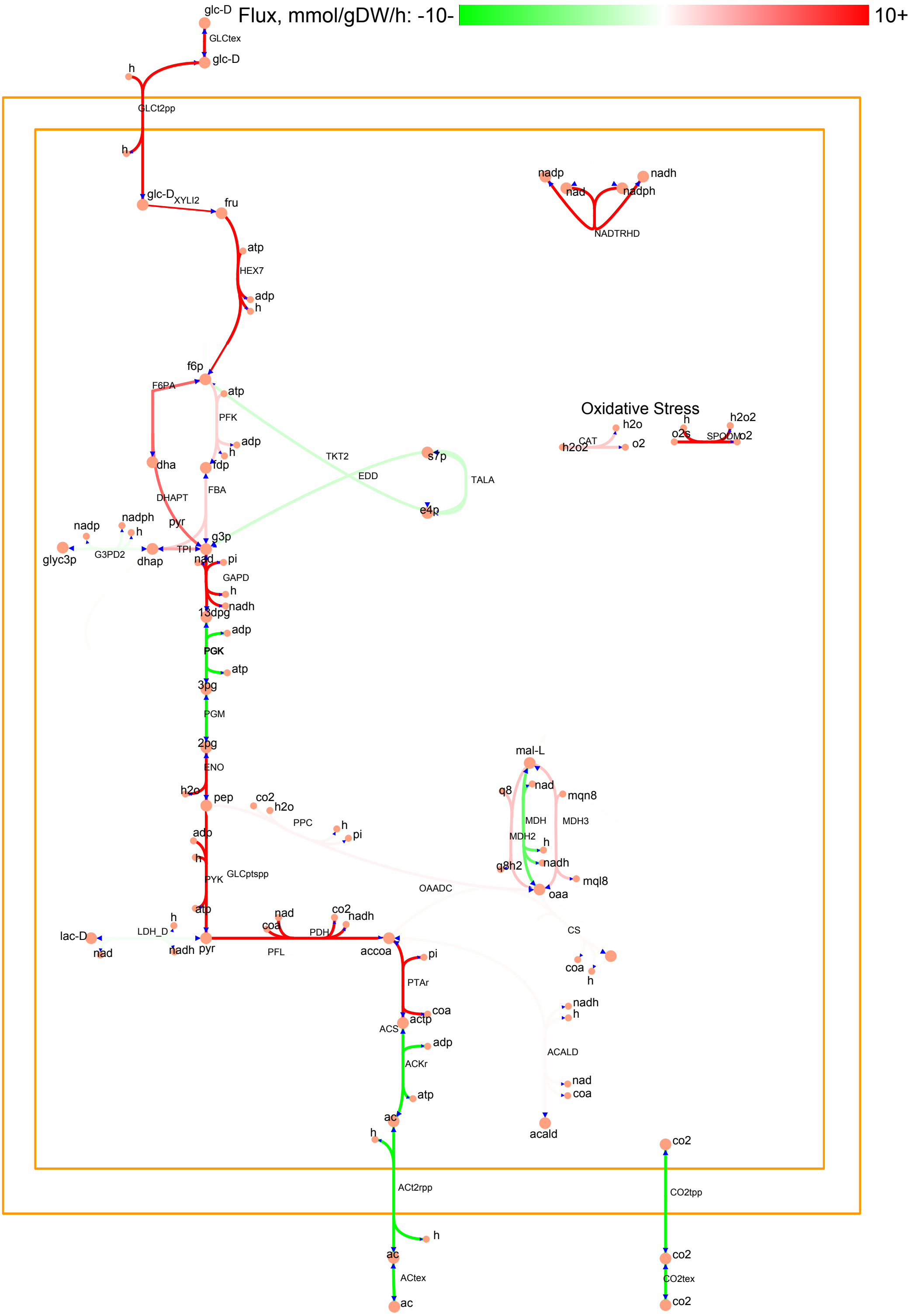

Figure 4.19: MUTE average core metabolism flux predictions for 5.5% salt. Without strict data filtering.

The difference between the data filtering regimes is already apparent in Figure 4.15. Glucose enters into glycolysis through the EMP pathway, eventually ending up as as pyruvate. As has been the case for all other predictions, there is high activity in the pathway that exports acetate, indicating overflow metabolism. There is high activity in parts of the Citric Acid Cycle, mostly maintained by internal loops in the metabolic network which artificially inflate flux strenghts. Notice that without strict data filtering, MUTE predicts high flux through reactions that scavenge superoxide radicals from the cytoplasm, a strong indicator that the cell is experiencing oxidative stress. Keeping in mind the established link between oxidative and osmotic stress, this could be an indication that including more data, although less accurate, improves the predictive power of the MUTE method.

Increasing the salt concentration to 3.5 % results in Figure 4.16 in significant predicted changes to the central metabolism. As for 2.0 % salt, glucose enters the EMP pathway of glycolysis, but an additional import of gluconate activates the ED pathway. The entry point into the ED pathway is downstream of the NADPH producing reactions, eliminating NADPH production as the motivation for the metabolic change. One difference between the EMP and ED pathways is that the ED pathway consumes slightly less ATP to produce g3p. This could be beneficial to the cell, but it does not explain why there is activity in glucose import reactions which bypass the ED pathway completely. The reactions responsible for handling superoxide radicals continue to carry high flux, and export of acetate seems to be the main carbon sink, incidating a persistance of oxidative stress.

When examining Figure 4.12, it appears that the glucose import into the cell is decreased. As was seen earlier for MUTE models with strict data filtering, this affects all downstream reactions, lowering their flux. Most reactions in the ED pathway are now only faintly visible, carrying what little flux is left into the same reactions as in 3.5 % salt. Once glycolysis is complete, pyruvate is converted and exported through the acetate export pathway, but there are two new options as well: D-lactate and pyruvate export. However, these two export reactions carry so little flux that it is hard for the naked eye to see, and the implications they bring are probably insignificant. Keep in mind, from section 4.3, that the expression levels at 4.5 % salt were lower than for the rest of the gene expression sets, resulting in fewer proteins being produced. The low flux on display in Figure 4.17 is likely a direct consequence of this, making it difficult to judge the accuracy of the predictions when compared to the other concentrations.

Continuing the increase in salt concentration, and by extension the severity of osmotic stress, Figure 4.13, representing 5.0 % salt, reveals some big changes from 4.5 % salt. The overall flux strength of the reactions is on average higher than at 4.5 % salt, strengthening the conjecture that low fluxes at 4.5 % salt are caused by low overall gene expression. The ED pathway is almost completely inactive, except for one reaction, which catalyzes the conversion of f6p and D-erythrose-4-phosphate (e4p) into sedoheptulose-7-phosphate (s7p) and g3p. By doing this, the cell is able to partially bypass the ATP investment required by the EMP pathway for catabolizing f6p into g3p and dhap. Downstream of g3p, after glycolysis is completed, the process of exporting acetate continues. Combined with the high flux through the superoxide radical scavenging reactions, this indicates that the cell continues to exhibit signs of oxidative stress.

At the highest salt concentration – 5.5 %, Figure 4.19 is very similar to Figure 4.18. This is in line with the results from MADE and the MUTE results without strict data filtering, which also had few differences between the two highest salt concentrations. This could be an indication that osmotic stress has a "sigmoidal" response curve, where osmoadaptations are at their maximum activity at approximately 5.0 % salt, after which the response to osmotic stress is unchanging.

Comparing the MUTE predictions without strict data filtering to those from MADE, it becomes obvious how different the two methods are. MADE quickly shuts down reactions in a discrete, binary-like fashion as it transitions between salt concentrations, while MUTE's predictions present a more continous and subtle change of metabolism. It is disappointing that MUTE failed to mirror the predictions by MADE of ATP synthase reversing in directionality, as this was perhaps the most interesting prediction produced.

4.4.1 Principal component analysis of sample data

The samples generated from the MUTE models with strict data filtering were subjected to PCA. A sorted list of PCA weights and their associated reactions is shown in table 4.3.

Table 4.3: Composition of first and second principal component for MUTE model samples without strict filtering.

First principal component

Reaction name	Weight
Lysophospholipase L2 (2-acylglycerophosphoglycerol, n-C14:0)	0.99
hydroxypyruvate isomerase	0.05
Hydrogenase (Demethylmenaquinone-8: 2 protons) (periplasm)	0.03
L-histidine transport via diffusion (extracellular to periplasm)	0.02
mercury (Hg+2) transport via diffusion (extracellular to periplasm)	0.02
LPS heptose kinase II (LPS core synthesis)	0.01

Second principal component

Reaction name	Weight
homoserine kinase	0.87
Hydroxypyruvate reductase (NADPH)	0.23
Hydroxypyruvate reductase (NADH)	0.22
3-(3-hydroxyphenyl)propionate transport via proton symport, reversible (periplasm)	0.14
L-homoserine transport via diffusion (extracellular to periplasm)	0.13
homoserine dehydrogenase (NADPH)	0.12

In table 4.3, the “Lysophospholipase L2” reaction completely dominates the first principal component weights, while “homoserine kinase” dominates in the second principal component. None of the reactions in table 4.3 were found to have known roles in osmotic stress.

The sample points were transformed into the first and second principal component space, grouped by salt concentration and visualized in Figure 4.20.

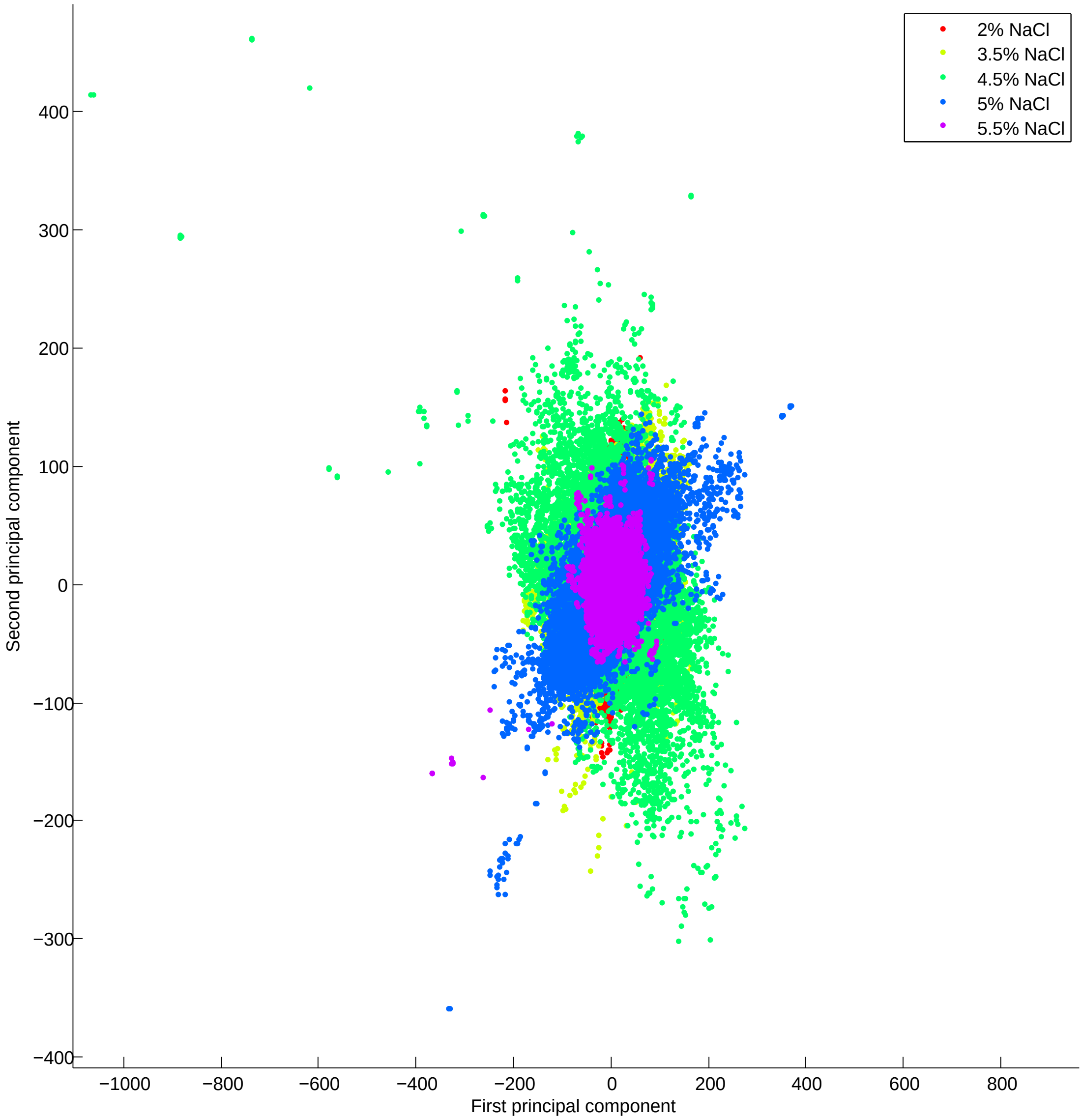

Figure 4.20: Scatterplot of PCA transformed sample points from all MUTE models with strict data filtering at the optimal growth surface.

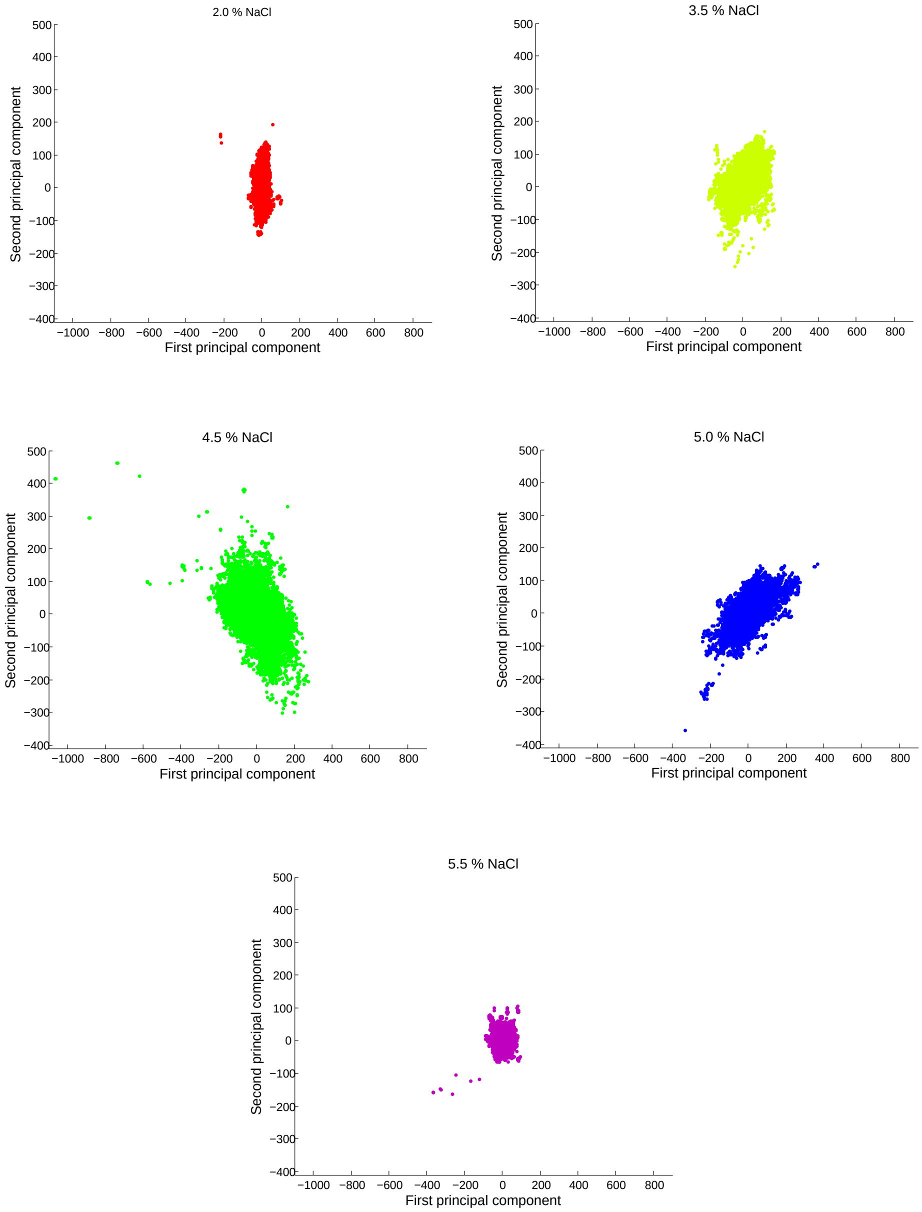

Figure 4.21: Separate PCA point clusters for each MUTE model with strict data filtering.

There is some separation of the different salt concentrations when viewing Figure 4.20, but they are mostly stacked on top of each other. In Figure 4.21 the salt concentration points have been separated from each other for easier visualization. Flux seems to be more tightly regulated at 5.5 % salt than for the other conditions, but there is no trend from 2.0 % to 5.0 % to make this meaningfull.

For the sampling of the MUTE models without strict data filtering, the components shown in table 4.4 were extracted from the first and second principal components. The PCA transformed sample points are visualized in Figures 4.22 and 4.23. The first principal component is almost completely dominated by an iron-sulphur cluster transport reaction, while the second component has more balanced weights. It is worth noting that SoxR, a protein responsible for activating an oxidative stress response in *E. coli*, is a homodimer with two [2Fe-2S] centers per dimer, possibly linking this reaction to the oxidative stress response predicted by both MADE and MUTE.[94]

The second principal component reaction who's weight is the largest in table 4.4, "gamma-butyrobetaine transport", transports gamma-butyrobetaine across the outer membrane of the cell and into the periplasm. This compound is a compatible osmoprotectant in *Listeria monocytogenes*, and could play a similar role in *E. coli.*[95] The remaining reactions in table 4.4 were not found to have connections to osmotic stress in *E. coli.*

Table 4.4: Composition of first and second principal component for MUTE model samples without strict filtering.

First principal component

Reaction name	Weight
ISC [2Fe-2S] Transfer	0.92
glycogen synthase (ADPGlc)	0.08
glycerate kinase	0.06
D-galacturonate transport via proton symport, reversible (periplasm)	0.06
Glycine Cleavage System	0.05
L-glutamate transport via proton symport, reversible (periplasm)	0.05
gamma-glutamylcysteine synthetase	0.05
glycolate transport via sodium symport (periplasm)	0.04
glucosyltransferase II (LPS core synthesis)	0.04
sn-glycerol-3-phosphoethanolamine transport via ABC system (periplasm)	0.04
D-glucarate transport via diffusion (extracellular to periplasm)	0.04
D-glycerate transport via diffusion (extracellular to periplasm)	0.04
D-galactose 1-phosphatase	0.04
Glycine betaine transport via ABC system (periplasm)	0.04
1,4-alpha-glucan branching enzyme (glycogen to bglycogen)	0.04

Second principal component

Reaction name	Weight
gamma-butyrobetaine transport via diffusion (extracellular to periplasm)	0.19
D-glucosamine transport via diffusion (extracellular to periplasm)	0.18
D-glucuronat transport via diffusion (extracellular to periplasm)	0.14
L-fucose transport via diffusion (extracellular to periplasm)	0.14
D-gluconate transport via proton symport, reversible (periplasm)	0.14
fumarate reductase	0.13
glutamate dehydrogenase (NADP)	0.13
ferric-dicitrate transport via ABC system (periplasm)	0.13
Fe-enterobactin reduction (Fe(III)-unloading)	0.12
glycerol-3-phosphate acyltransferase (C12:0)	0.11
Fructose transport via PEP:Pyr PTS (f6p generating) (periplasm)	0.11
Glucose-6-phosphate transport via phosphate antiport (periplasm)	0.11
guanylate kinase (GMP:ATP)	0.11
glutaminase	0.11
Fe-enterobactin transport via ton system (extracellular)	0.10

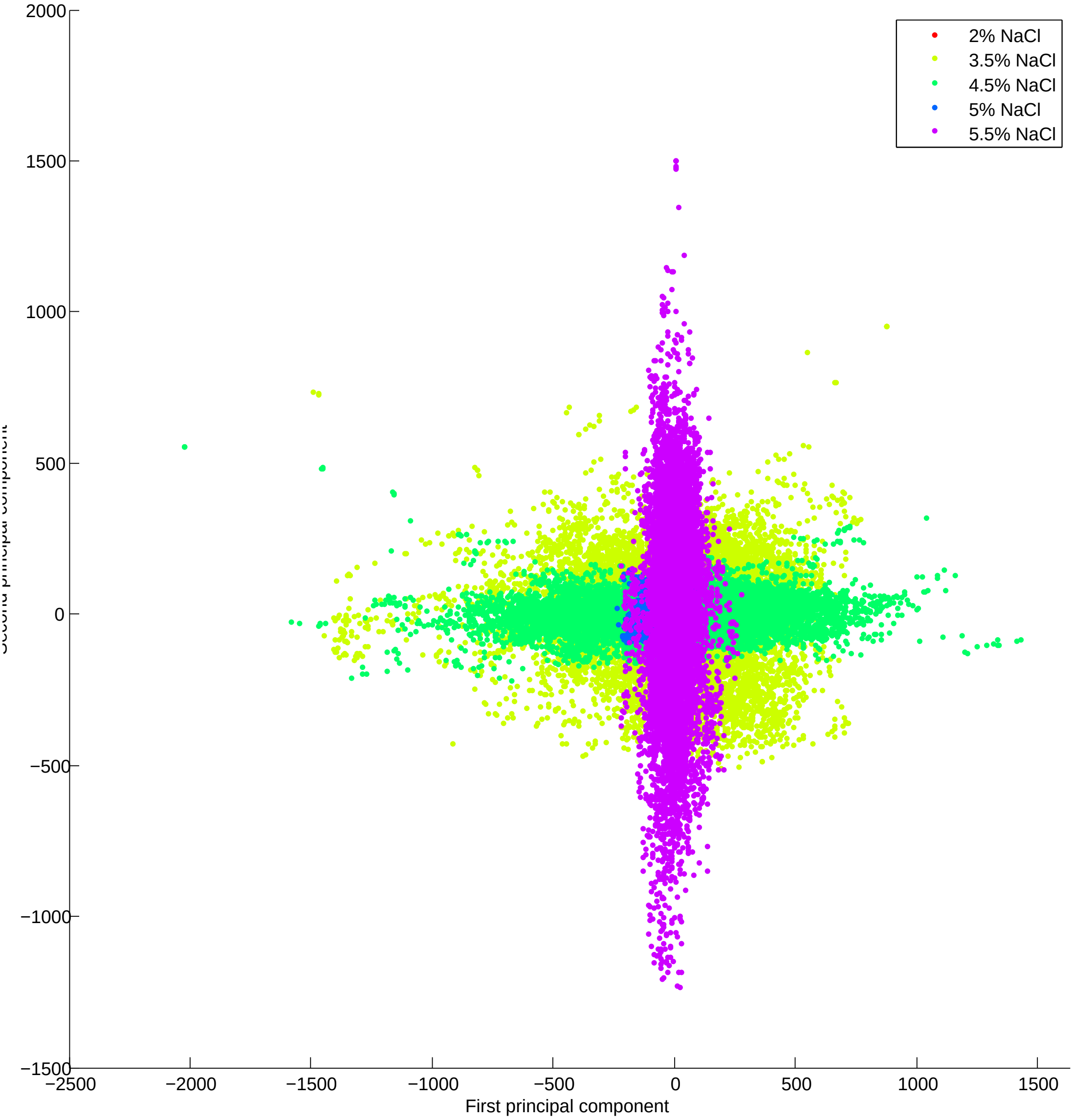

Figure 4.22: Scatterplot of PCA transformed sample points from all MUTE models without strict data filtering at the optimal growth surface.

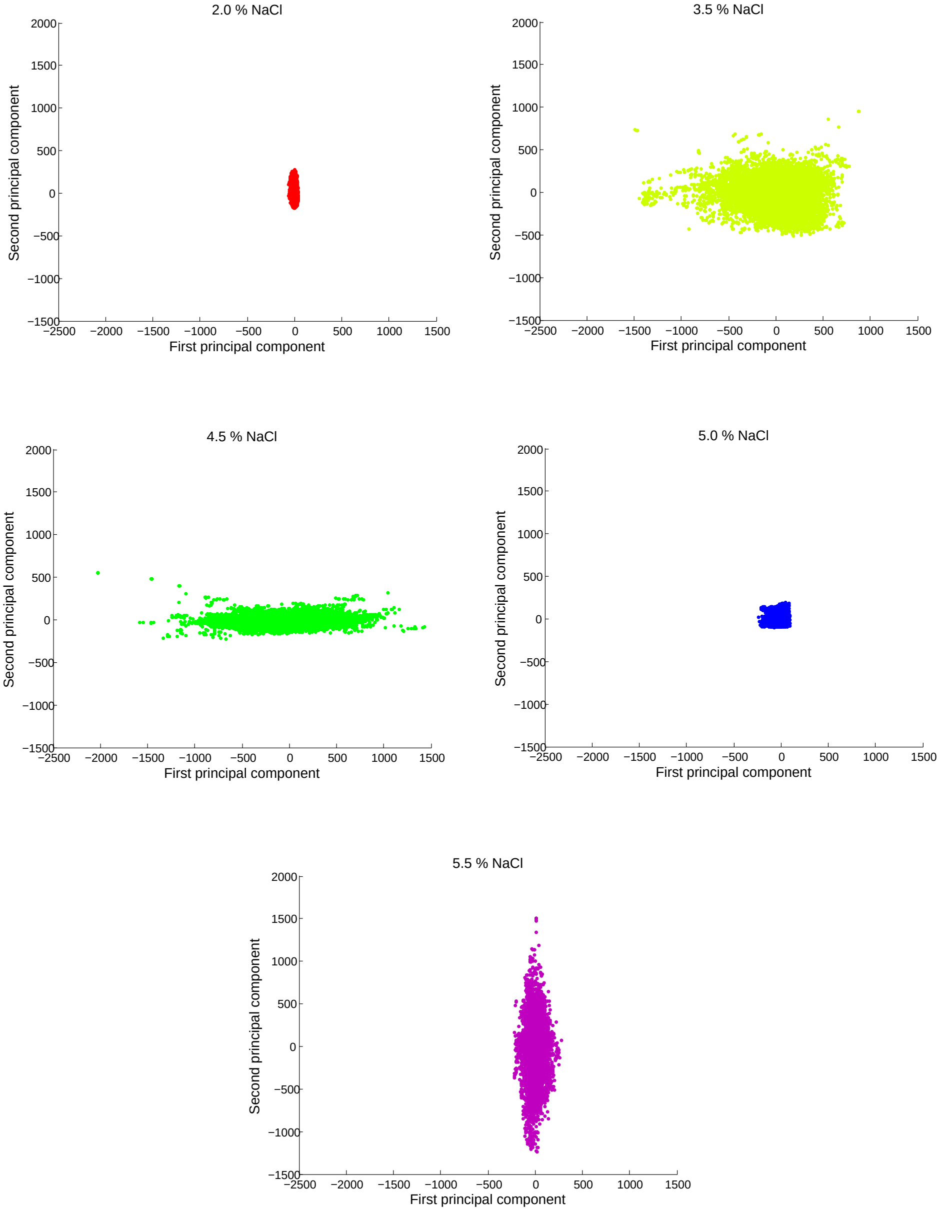

Figure 4.23: Separate PCA point clusters for each MUTE model without strict data filtering.

The failure of PCA to detect any reactions relevant for osmoadaption in the sample data from the strict filtering regime of MUTE could be an indication that the PCA method is poorly suited for investigating stress. As PCA identifies reactions with high variability, it would miss reactions that are tightly regulated, as might be the case for stress response reactions during osmoadaptation.

An alternative approach to PCA was attempted, where reactions with low flux variance in each sample, but where the mean flux of those reactions varied greatly between the samples, were identified and investigated. This resulted in a list of reactions who's sample point distribution resembled tight clusters distant from each other. Table 4.5 shows a list of the top five reactions fulfilling these criteria for each filtering regime.

Table 4.5: Reactions who's points group into distinct clusters grouped by salt concentration.

MUTE with strict data filtering

Reaction name	Variance of means, $\frac{\text{mmol}}{\text{gDW} \cdot \text{h}}$
isochorismate synthase	35
phosphoribosylpyrophosphate synthetase	24
uracil transport in via proton symport	23
cytidine transport in via proton symport	22
phosphopentomutase	21

MUTE without strict data filtering

Reaction name	Variance of means, $\frac{\text{mmol}}{\text{gDW} \cdot \text{h}}$
dihydropteridine reductase	154
L-idonate 5-dehydrogenase	31
thioredoxin reductase (NADPH)	25
uracil transport in via proton symport	19
L-valine reversible transport via proton symport	16

None of the reactions listed in table 4.5 were associated with stress responses in *E. coli*, except thioredoxin reductase which is involved in maintaining a reduced environment in the cytoplasm, connecting it to oxidative stress.[94]

Chapter 5

Conclusion

This thesis has achieved its' primary goal of modelling osmotic stress in *E. coli*. The newly developed method for integrating gene expression data and metabolic models, MUTE, was able to predict realistic metabolic fluxes constraints based on predicted protein concentrations.

Investigating osmotic stress using both the MADE and MUTE methods resulted in predictions that implicated overflow metabolism in *E. coli* cells during osmotic stress, indicating that oxidative stress adaptation is activated during osmoadaptation.

At 5.0 % and 5.5 % salt, MADE predicted that ATP synthase runs in reverse, transporting protons across the cell membrane from the cytoplasm into the periplasm, at the expense of ATP. At these two salt concentrations, MADE predicted no biomass production for *E. coli*, possibly hinting at some stationary phase phenotype focused on surviving.

MUTE models with strict data filtering seemed unable to predict significant changes in the core metabolism of *E. coli* during osmotic stress, likely caused by filtering away important experimental data. The predictions became more responsive to changing conditions once less stringent data filtering regimes were employed, suggesting that more data with less accuracy is better than the opposite.

Comparisons of the MADE and MUTE methods showed that MUTE maintained metabolic network connectivity and flux through several salt concentration transitions, while MADE's binary interpretation of gene expression data quickly disabled entire pathways.

Sampling and analyzing MUTE models by PCA revealed high variance in the activity of the gamma-butyrobetaine transport reaction. Previously reported as a compatible osmoprotectant in *L. monocytogenes*, import of gamma-butyrobetaine could be an important part of osmoadaptation in *E. coli*.

Further research into osmotic stress in *E. coli* should investigate the presence of gamma-butyrobetaine in the cytosol during osmotic stress. Measuring the pH of *E. coli*'s periplasm at high salt concentrations could be done using pH sensitive fluorescent proteins in order to investigate the prediction that ATP synthase is run in reverse in these conditions.

Appendix A

Visualization of MUTE flux constraints (with strict data filtering)

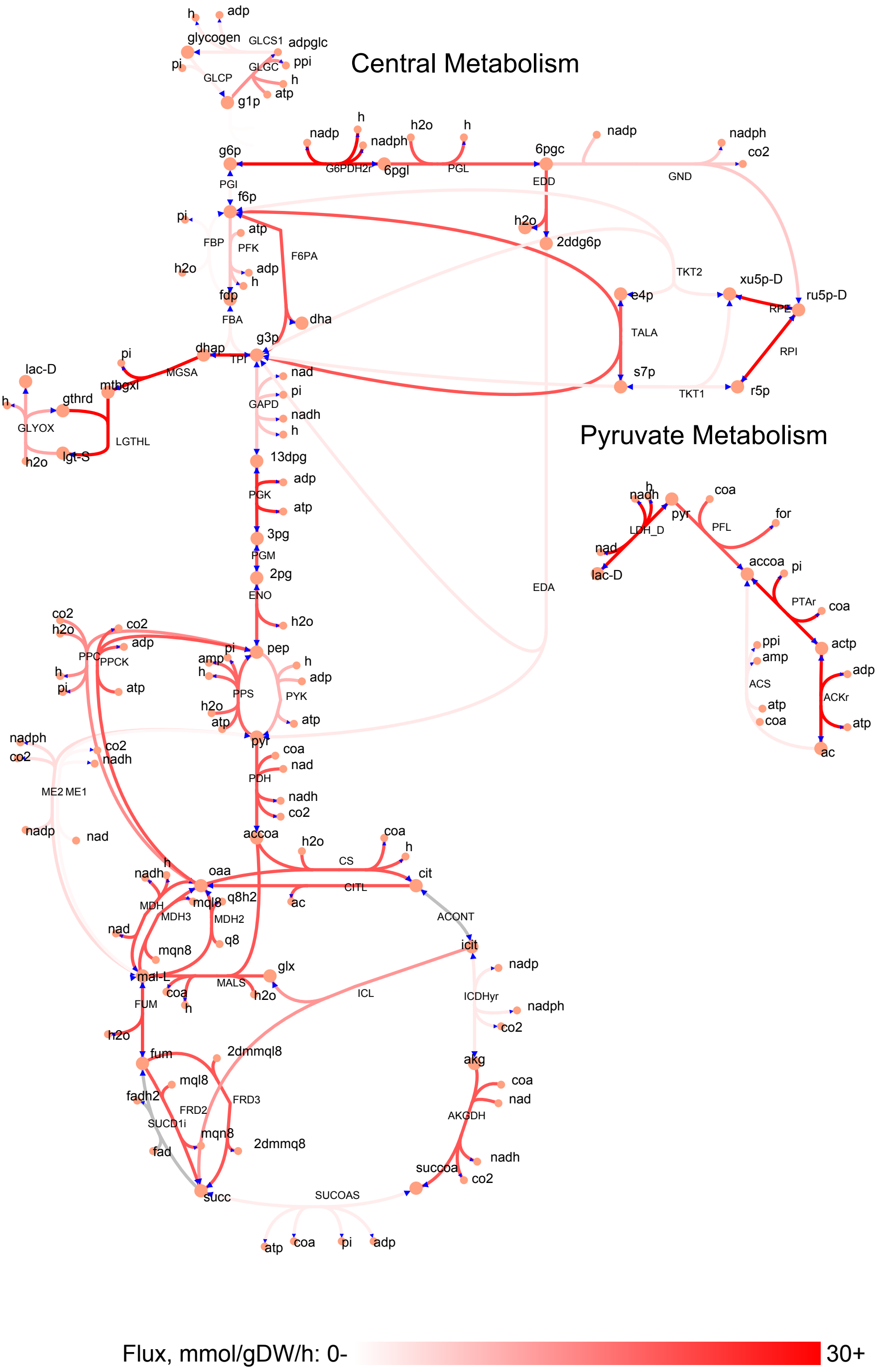

Figure A.0.1: MUTE constraints for central metabolism at 2 % NaCl. Model generated with strict filtering.

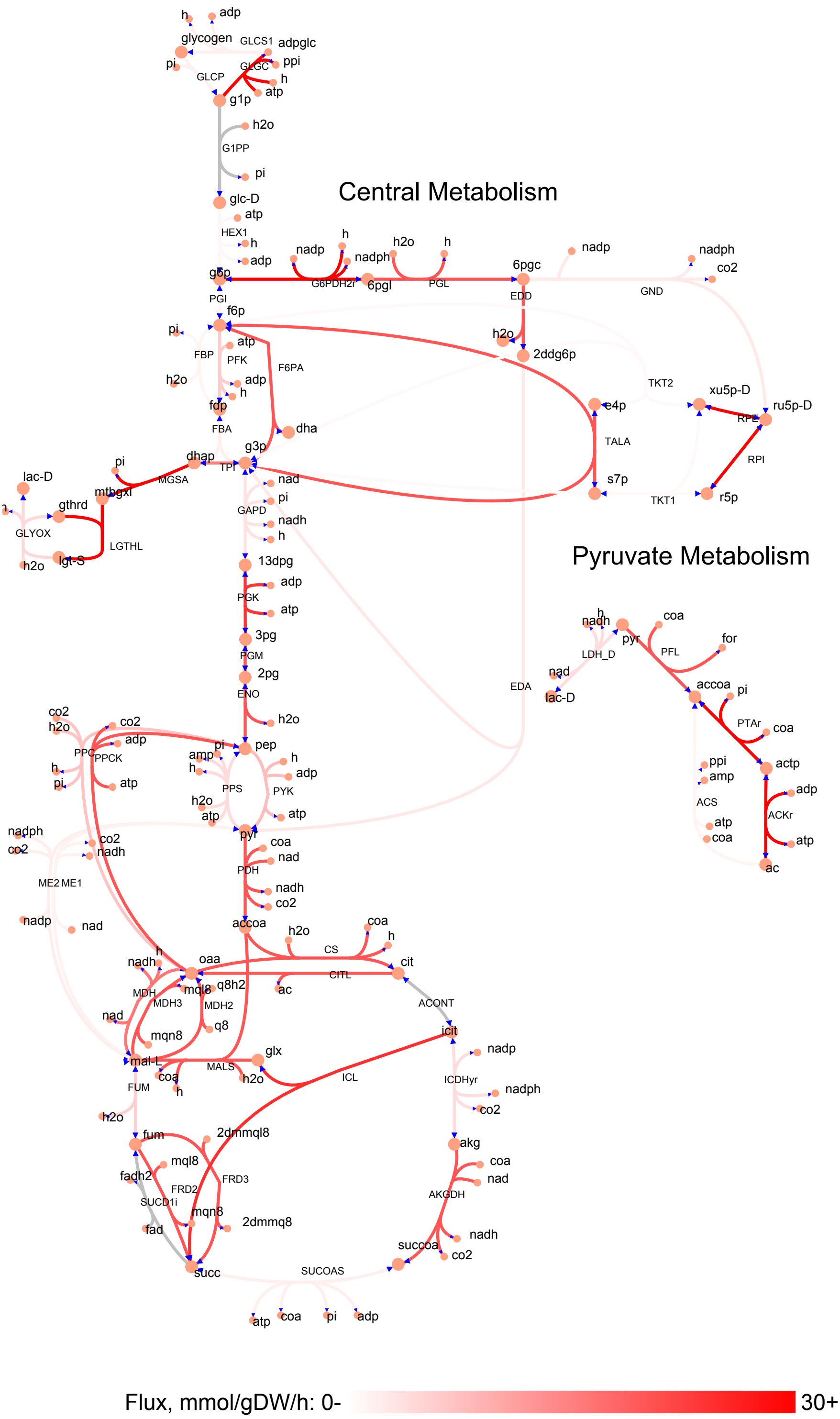

Figure A.0.2: MUTE constraints for central metabolism at 3.5 % NaCl. Model generated with strict filtering.

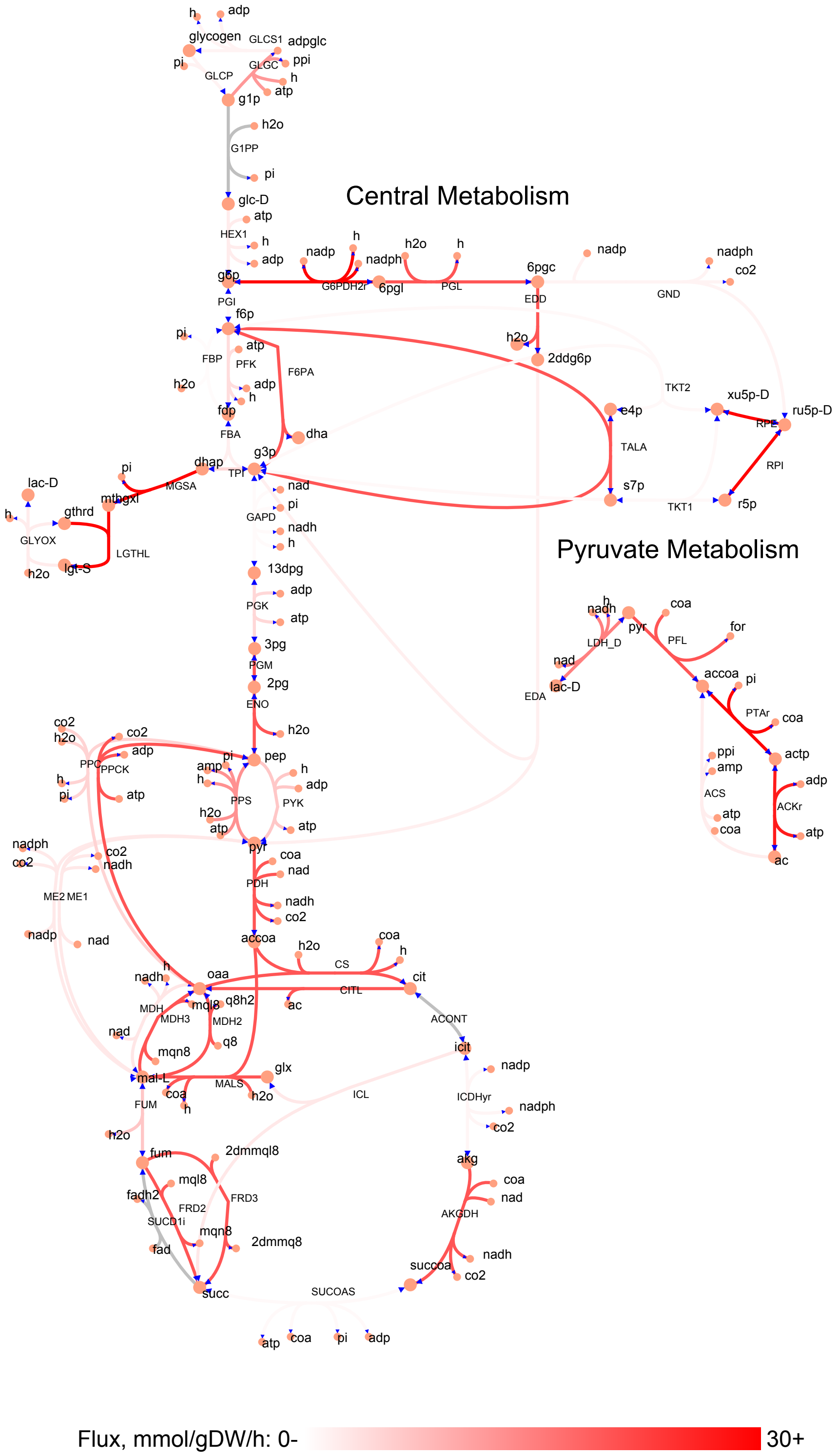

Figure A.0.3: MUTE constraints for central metabolism at 4.5 % NaCl. Model generated with strict filtering.

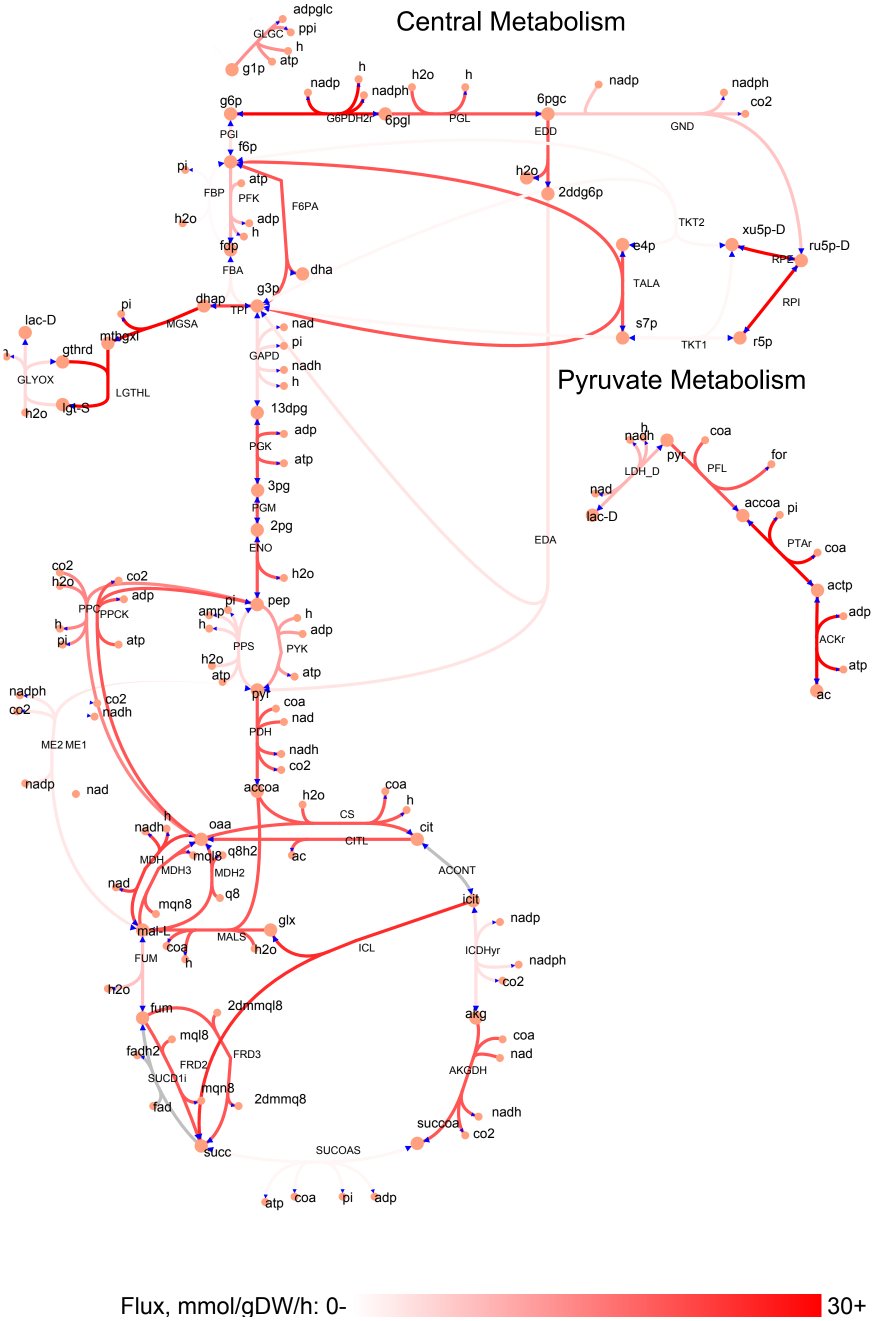

Figure A.0.4: MUTE constraints for central metabolism at 5 % NaCl. Model generated with strict filtering.

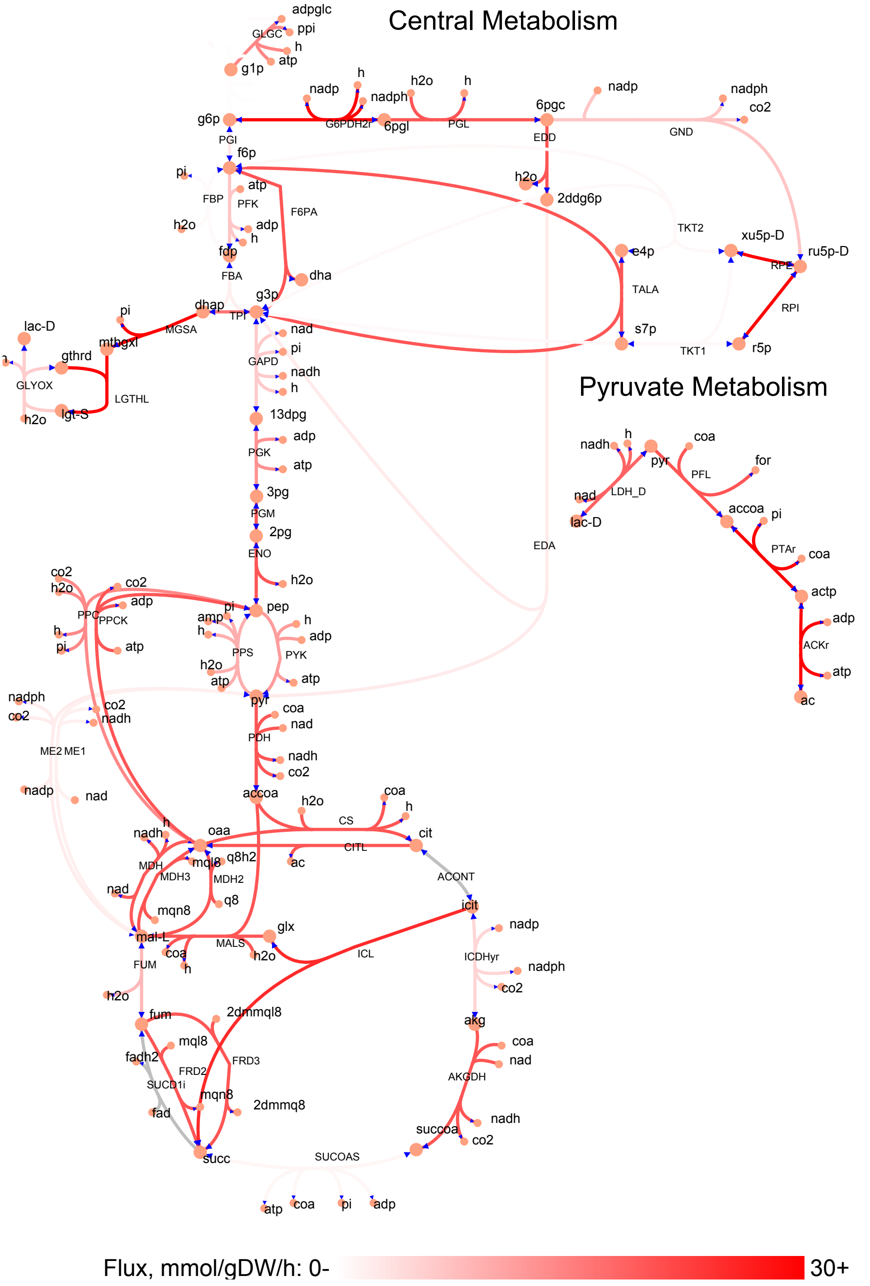

Figure A.0.5: MUTE constraints for central metabolism at 5.5 % NaCl. Model generated with strict filtering.

Appendix B

Visualization of MUTE flux constraints (without strict data filtering)

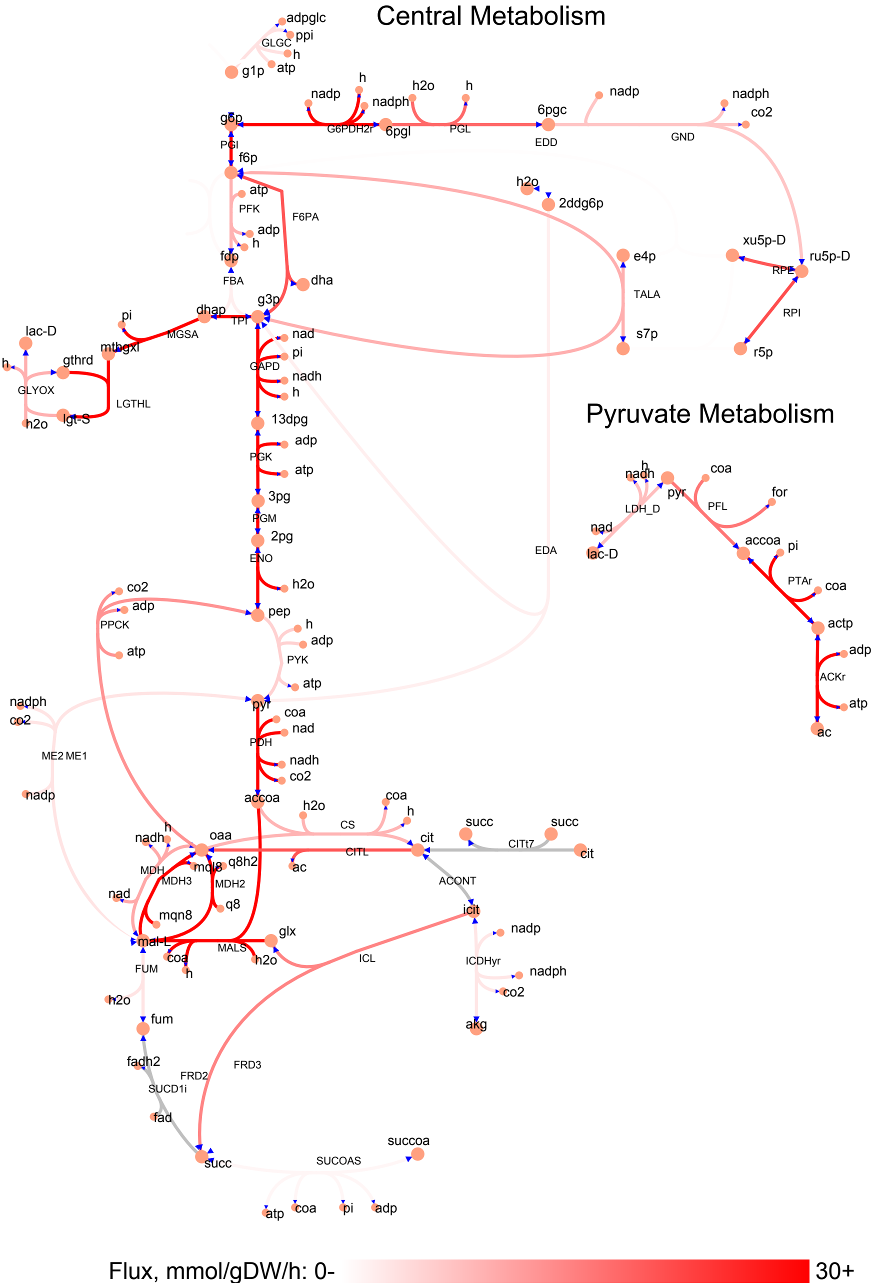

Figure B.0.1: MUTE constraints for central and pyruvate metabolism at 2 % NaCl. Model generated without strict filtering.

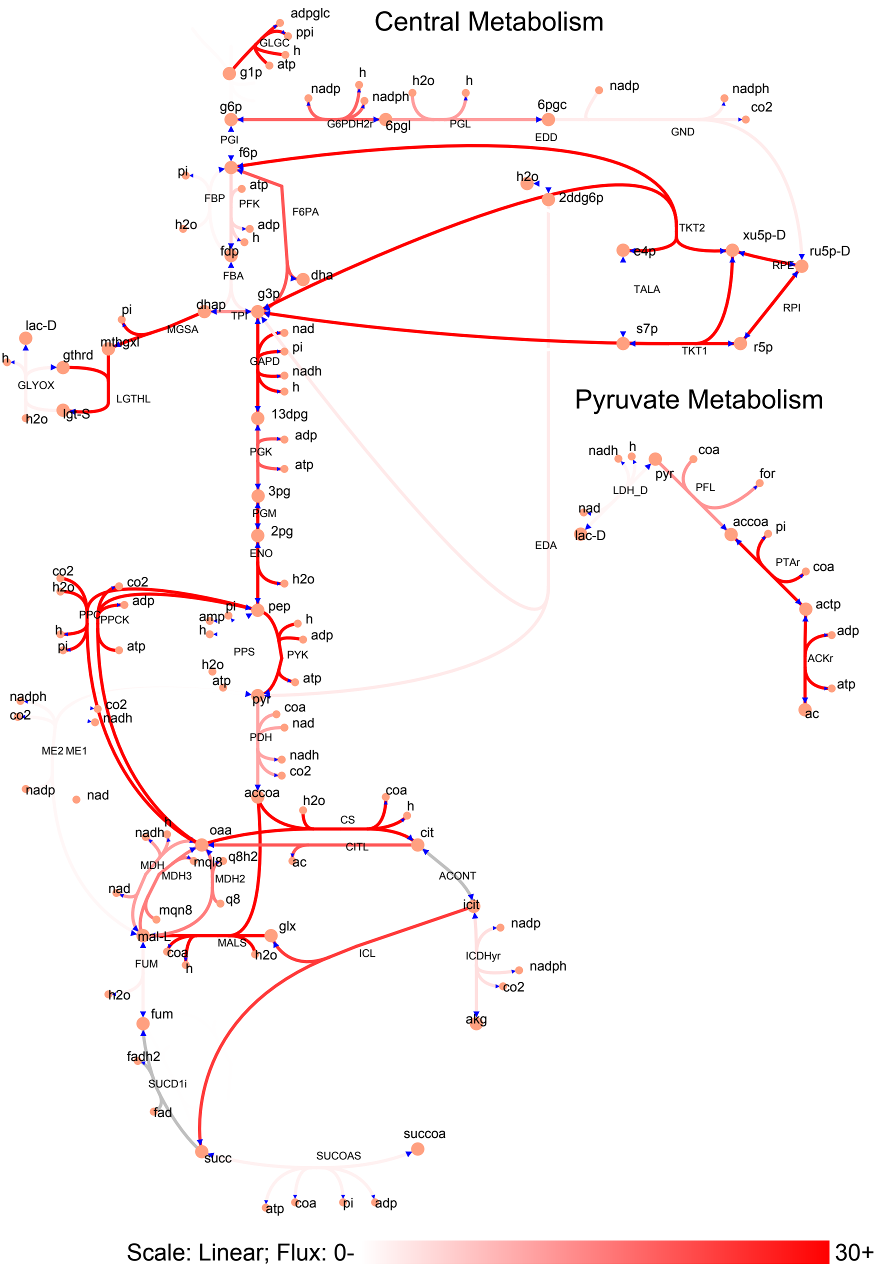

Figure B.0.2: MUTE constraints for central and pyruvate metabolism at 3.5 % NaCl. Model generated without strict filtering.

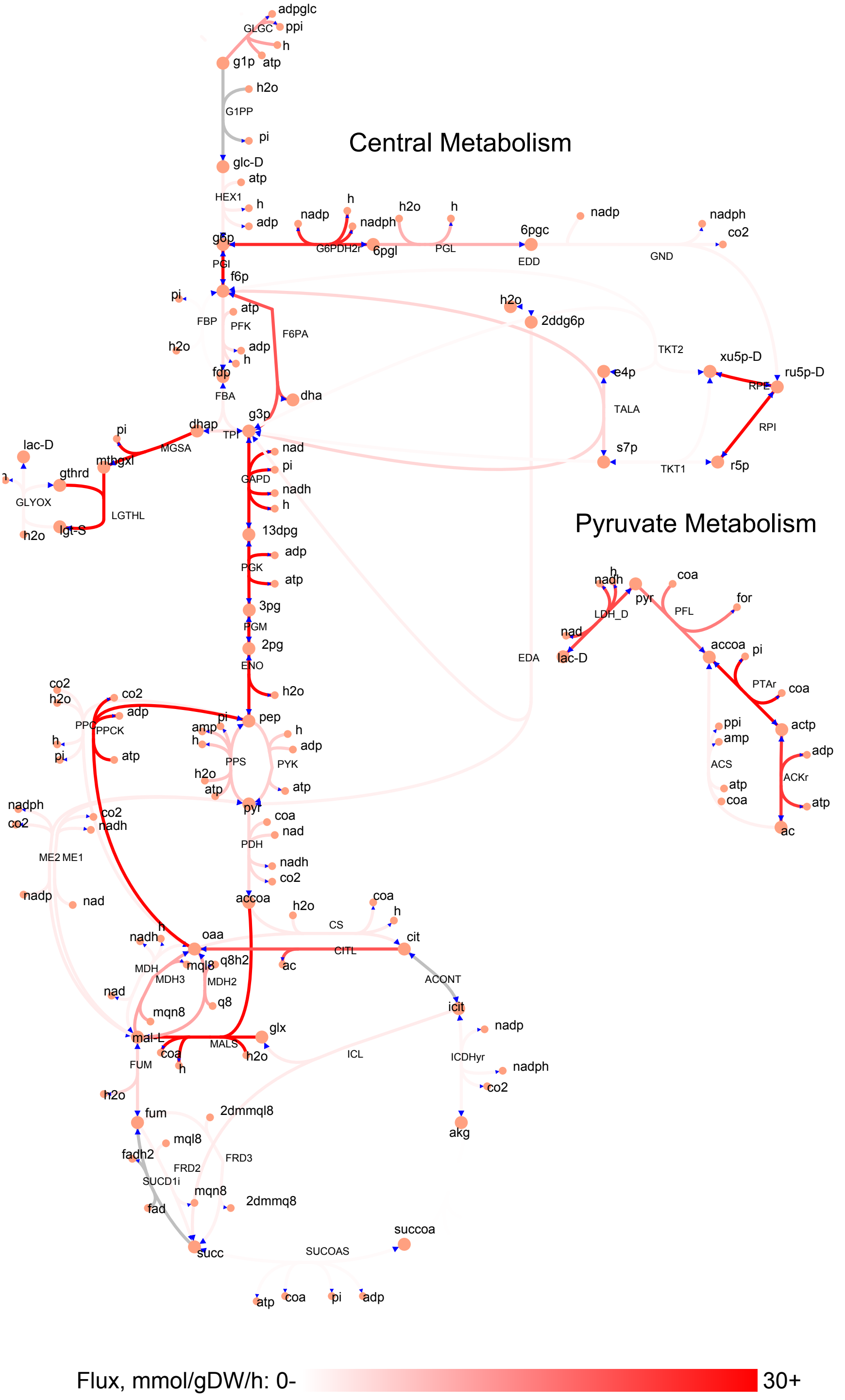

Figure B.0.3: MUTE constraints for central and pyruvate metabolism at 4.5 % NaCl. Model generated without strict filtering.

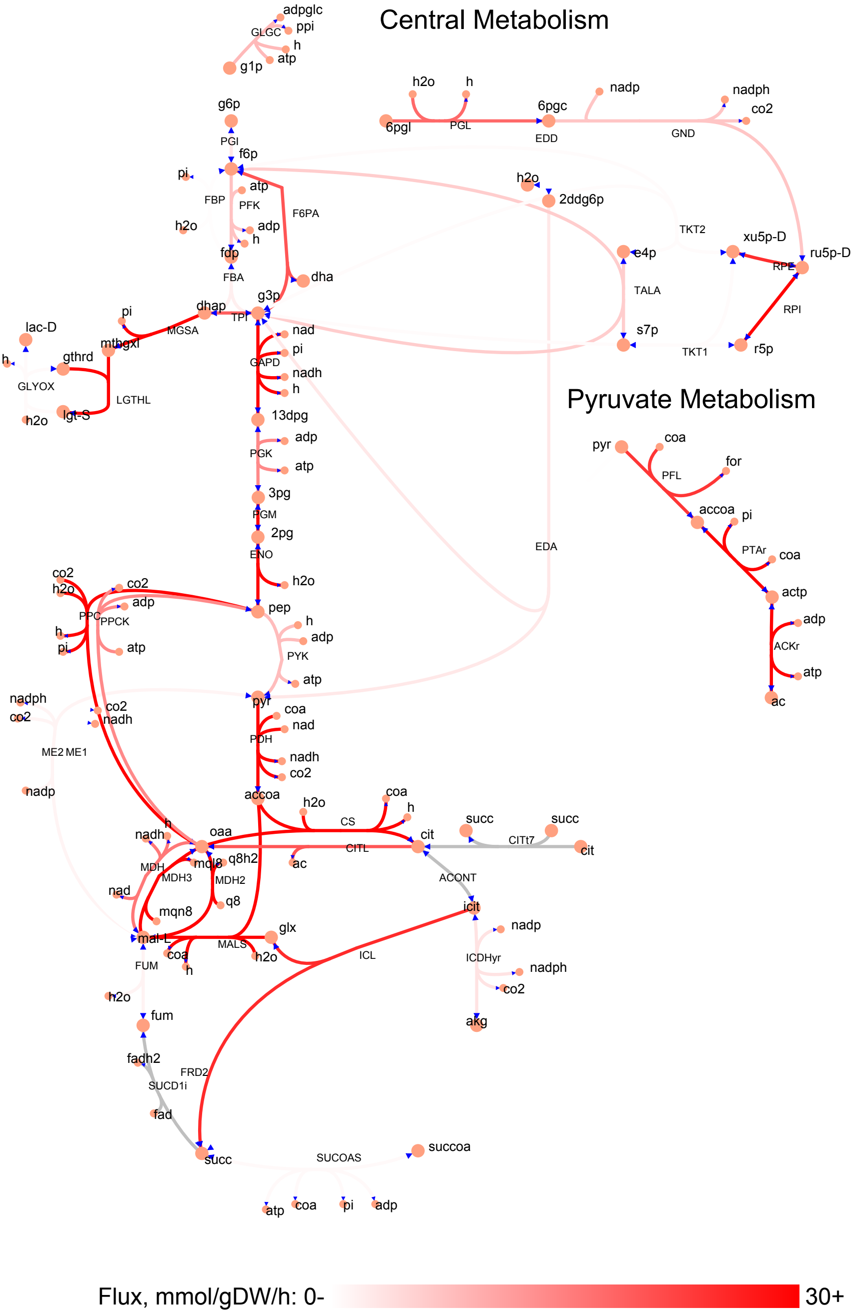

Figure B.0.4: MUTE constraints for central and pyruvate metabolism at 5 % NaCl. Model generated without strict filtering.

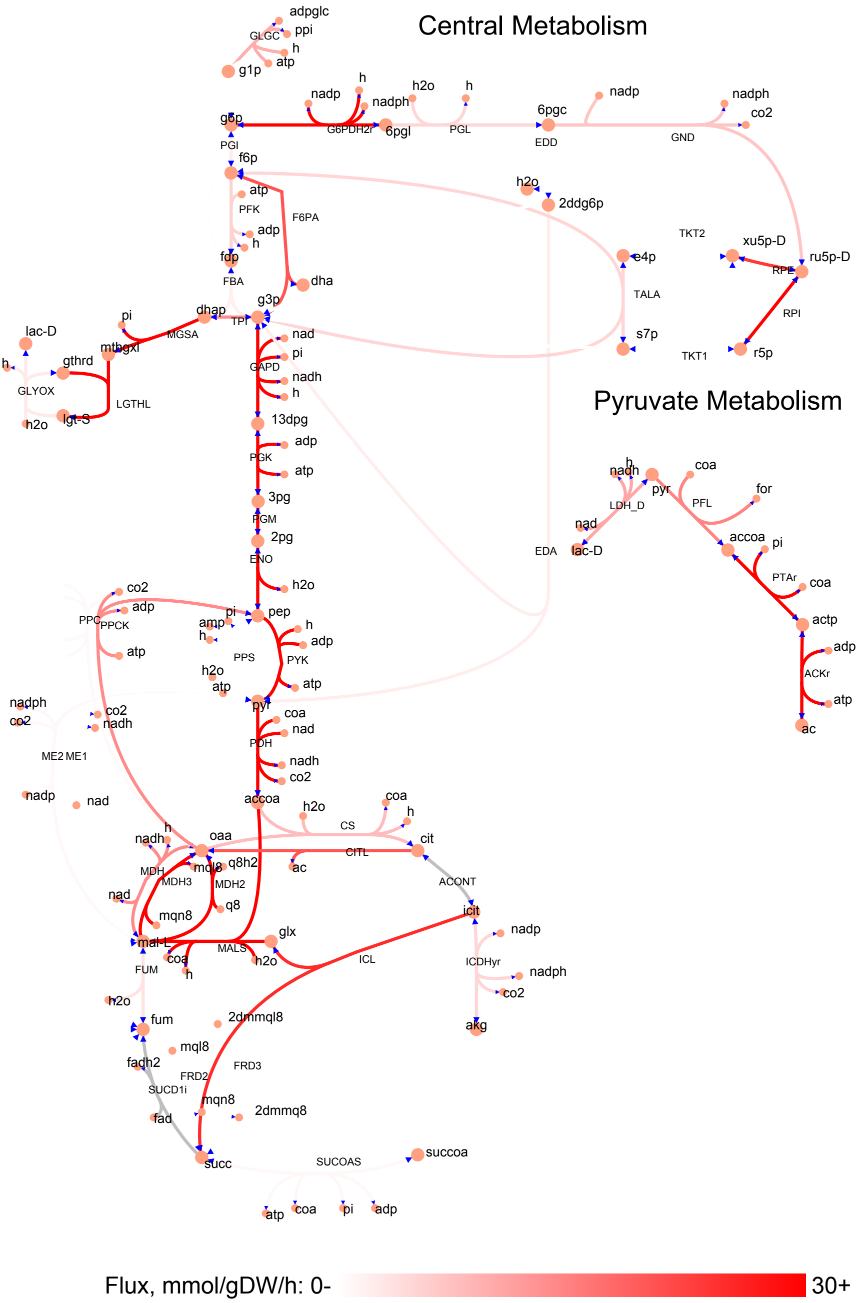

Figure B.0.5: MUTE constraints for central and pyruvate metabolism at 5.5 % NaCl. Model generated without strict filtering.

Bibliography

[1] Jeremy E Purvis, LP Yomano, and LO Ingram. Enhanced trehalose production improves growth of escherichia coli under osmotic stress. *Applied and environmental microbiology*, 71(7):3761–3769, 2005.

[2] Laszlo N Csonka. Physiological and genetic responses of bacteria to osmotic stress. *Microbiological reviews*, 53(1):121–147, 1989.

[3] A Metris, SM George, F Mulholland, AT Carter, and J Baranyi. Metabolic shift of escherichia coli under salt stress in the presence of glycine betaine. *Applied and environmental microbiology*, 80(15):4745–4756, 2014.

[4] Sue Shephard. *Pickled, Potted, and Canned: How the Art and Science of Food Preserving Changed the World*. Simon and Schuster, 2006.

[5] Ethan B Solomon, Sima Yaron, and Karl R Matthews. Transmission of escherichia coli o157: H7 from contaminated manure and irrigation water to lettuce plant tissue and its subsequent internalization. *Applied and Environmental Microbiology*, 68(1):397–400, 2002.

[6] James B. Kaper, James P. Nataro, and Harry L. T. Mobley. Pathogenic escherichia coli. *Nature Reviews Microbiology*, 2(2):123–140, Feb 2004.

[7] Multistate outbreak of shiga toxin-producing escherichia coli o121 infections linked to raw clover sprouts. Online, 2014.

[8] Multistate outbreak of shiga toxin-producing escherichia coli o157:h7 infections linked to ground beef. Online, 2014.

[9] Firdausi Qadri, Ann-Mari Svennerholm, ASG Faruque, and R Bradley Sack. Enterotoxigenic escherichia coli in developing countries: epidemiology, microbiology, clinical features, treatment, and prevention. *Clinical microbiology reviews*, 18(3):465–483, 2005.

[10] Yasushi Ishihama, Thorsten Schmidt, Juri Rappsilber, Matthias Mann, F Ulrich Hartl, Michael J Kerner, and Dmitrij Frishman. Protein abundance profiling of the escherichia coli cytosol. *BMC genomics*, 9(1):102, 2008.

[11] Albert-László Barabási and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.

[12] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. The sequence of the human genome. *science*, 291(5507):1304–1351, 2001.

[13] Michael L Metzker. Sequencing technologies—the next generation. *Nature Reviews Genetics*, 11(1):31–46, 2009.

[14] Elaine R Mardis. The impact of next-generation sequencing technology on genetics. *Trends in genetics*, 24(3):133–141, 2008.

[15] Jay Shendure and Hanlee Ji. Next-generation dna sequencing. *Nature biotechnology*, 26(10):1135–1145, 2008.

[16] Upinder S Bhalla and Ravi Iyengar. Emergent properties of networks of biological signaling pathways. *Science*, 283(5400):381–387, 1999.

[17] Benjamin B. Machta, Ricky Chachra, Mark K. Transtrum, and James P. Sethna. Parameter space compression underlies emergent theories and predictive models. *Science*, 342(6158):604–607, 2013.

[18] Stanley Wasserman. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[19] David Sprinzak and Michael B Elowitz. Reconstruction of genetic circuits. *Nature*, 438(7067):443–448, 2005.

[20] G Joshi-Tope, Marc Gillespie, Imre Vastrik, Peter D'Eustachio, Esther Schmidt, Bernard de Bono, Bijay Jassal, GR Gopinath, GR Wu, Lisa Matthews, et al. Reactome: a knowledgebase of biological pathways. *Nucleic acids research*, 33(suppl 1):D428–D432, 2005.

[21] Ines Thiele and Bernhard Ø Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols*, 5(1):93–121, 2010.

[22] Markus W Covert and Bernhard O Palsson. Constraints-based models: regulation of gene expression reduces the steady-state solution space. *Journal of theoretical biology*, 221(3):309–325, 2003.

[23] Ali Navid and Eivind Almaas. Genome-level transcription data of yersinia pestis analyzed with a new metabolic constraint-based approach. *BMC systems biology*, 6(1):150, 2012.

[24] Jennifer L Reed. Shrinking the metabolic solution space using experimental datasets. *PLoS computational biology*, 8(8):e1002662, 2012.

[25] Paul A Jensen and Jason A Papin. Functional integration of a metabolic network model and expression data without arbitrary thresholding. *Bioinformatics*, 27(4):541–547, 2011.

[26] Douglas B Kell. Systems biology, metabolic modelling and metabolomics in drug discovery and development. *Drug discovery today*, 11(23):1085–1092, 2006.

[27] Stanley Brul, Suzanne Van Gerwen, and Marcel Zwietering. *Modelling microorganisms in food*. Elsevier, 2007.

[28] Ori Folger, Livnat Jerby, Christian Frezza, Eyal Gottlieb, Eytan Ruppin, and Tomer Shlomi. Predicting selective drug targets in cancer through metabolic networks. *Molecular systems biology*, 7(1):501, 2011.

[29] H Kruse. Globalization of the food supply–food safety implications: Special regional requirements: future concerns. *Food control*, 10(4):315–320, 1999.

[30] N Beales. Adaptation of microorganisms to cold temperatures, weak acid preservatives, low ph, and osmotic stress: a review. *Comprehensive Reviews in Food science and Food safety*, 3(1):1–20, 2004.

[31] Dov Greenbaum, Christopher Colangelo, Kenneth Williams, and Mark Gerstein. Comparing protein abundance and mrna expression levels on a genomic scale. *Genome Biol*, 4(9):117, 2003.

[32] Nicholas T Ingolia, Sina Ghaemmaghami, John RS Newman, and Jonathan S Weissman. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *science*, 324(5924):218–223, 2009.

[33] Gene-Wei Li, David Burkhardt, Carol Gross, and Jonathan S Weissman. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell*, 157(3):624–635, 2014.

[34] F. R. Blattner. The complete genome sequence of escherichia coli k-12. *Science*, 277(5331):1453–1462, Sep 1997.

[35] Johanna Roostalu, Arvi Jõers, Hannes Luidalepp, Niilo Kaldalu, and Tanel Tenson. Cell division in escherichia coli cultures monitored at single cell resolution. *BMC Microbiology*, 8(1):68, 2008.

[36] Ingrid M Keseler, Julio Collado-Vides, Alberto Santos-Zavaleta, Martin Peralta-Gil, Socorro Gama-Castro, Luis Muñiz-Rascado, César Bonavides-Martinez, Suzanne Paley, Markus Krummenacker, Tomer Altman, et al. Ecocyc: a comprehensive database of escherichia coli biology. *Nucleic acids research*, 39(suppl 1):D583–D590, 2011.

[37] Bjarne Landfald and Arne R Strøm. Choline-glycine betaine pathway confers a high level of osmotic tolerance in escherichia coli. *Journal of bacteriology*, 165(3):849–855, 1986.

[38] Lindsay Murdock, Tangi Burke, Chelsea Coumoundouros, Doreen E Culham, Charles E Deutch, James Ellinger, Craig H Kerr, Samantha M Plater, Eric To, Geordie Wright, et al. Analysis of strains lacking known osmolyte accumulation mechanisms reveals contributions of osmolytes and transporters to protection against abiotic stress. *Applied and environmental microbiology*, 80(17):5366–5378, 2014.

[39] Montserrat Argandoña, Joaquín J Nieto, Fernando Iglesias-Guerra, Maria Isabel Calderón, Raúl García-Estepa, and Carmen Vargas. Interplay between iron homeostasis and the osmotic stress response in the halophilic bacterium chromohalobacter salexigens. *Applied and environmental microbiology*, 76(11):3575–3589, 2010.

[40] Hanan Gancz and D Scott Merrell. The helicobacter pylori ferric uptake regulator (fur) is essential for growth under sodium chloride stress. *The Journal of Microbiology*, 49(2):294–298, 2011.

[41] Tamara Hoffmann, Alexandra Schütz, Margot Brosius, Andrea Völker, Uwe Völker, and Erhard Bremer. High-salinity-induced iron limitation in bacillus subtilis. *Journal of bacteriology*, 184(3):718–727, 2002.

[42] Aline Metris, Susan George, and József Baranyi. Modelling osmotic stress by flux balance analysis at the genomic scale. *International journal of food microbiology*, 152(3):123–128, 2012.

[43] Jeffrey D Orth, Ines Thiele, and Bernhard Ø Palsson. What is flux balance analysis? *Nature biotechnology*, 28(3):245–248, 2010.

[44] Santiago Schnell. Validity of the michaelis–menten equation–steady-state or reactant stationary assumption: that is the question. *FEBS Journal*, 281(2):464–472, 2014.

[45] Nathan D Price, Jennifer L Reed, Jason A Papin, Sharon J Wiback, and Bernhard O Palsson. Network-based analysis of metabolic regulation in the human red blood cell. *Journal of Theoretical Biology*, 225(2):185–194, 2003.

[46] Cong T Trinh, Aaron Wlaschin, and Friedrich Srienc. Elementary mode analysis: a useful metabolic pathway analysis tool for characterizing cellular metabolism. *Applied microbiology and biotechnology*, 81(5):813–826, 2009.

[47] Abdelhalim Larhlimi and Alexander Bockmayr. A new constraint-based description of the steady-state flux cone of metabolic networks. *Discrete Applied Mathematics*, 157(10):2257–2266, 2009.

[48] Roi Adadi, Benjamin Volkmer, Ron Milo, Matthias Heinemann, and Tomer Shlomi. Prediction of microbial growth rate versus biomass yield by a metabolic network with kinetic parameters. *PLoS computational biology*, 8(7):e1002575, 2012.

[49] Jennifer L Reed, Thuy D Vo, Christophe H Schilling, Bernhard O Palsson, et al. An expanded genome-scale model of escherichia coli k-12 (ijr904 gsm/gpr). *Genome Biol*, 4(9):R54, 2003.

[50] Adam M Feist, Christopher S Henry, Jennifer L Reed, Markus Krummenacker, Andrew R Joyce, Peter D Karp, Linda J Broadbelt, Vassily Hatzimanikatis, and Bernhard A Palsson. A genome-scale metabolic reconstruction for escherichia coli k-12 mg1655 that accounts for 1260 ORfs and thermodynamic information. *Molecular Systems Biology*, 3, Jun 2007.

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

[52] Sarah M Keating, Benjamin J Bornstein, Andrew Finney, and Michael Hucka. Sbmltoolbox: an sbml toolbox for matlab users. *Bioinformatics*, 22(10):1275–1277, 2006.

[53] Scott A Becker, Adam M Feist, Monica L Mo, Gregory Hannum, Bernhard Ø Palsson, and Markus J Herrgard. Quantitative prediction of cellular metabolism with constraint-based models: the cobra toolbox. *Nature protocols*, 2(3):727–738, 2007.

[54] Reiner Horst and H Edwin Romeijn. *Quadratic Optimization*, volume 2. Springer Science & Business Media, 2002.

[55] Moritz Fleischmann, Jacqueline M Bloemhof-Ruwaard, Rommert Dekker, Erwin Van der Laan, Jo AEE Van Nunen, and Luk N Van Wassenhove. Quantitative models for reverse logistics: A review. *European journal of operational research*, 103(1):1–17, 1997.

[56] George B Dantzig, Alex Orden, Philip Wolfe, et al. The generalized simplex method for minimizing a linear form under linear inequality restraints. *Pacific Journal of Mathematics*, 5(2):183–195, 1955.

[57] Hoang Tuy. *Convex analysis and global optimization*, volume 22. Springer Science & Business Media, 1998.

[58] David L Applegate, William Cook, Sanjeeb Dash, and Daniel G Espinoza. Exact solutions to linear programming problems. *Operations Research Letters*, 35(6):693–699, 2007.

[59] Panos M Pardalos and Georg Schnitger. Checking local optimality in constrained quadratic programming is np-hard. *Operations Research Letters*, 7(1):33–35, 1988.

[60] A Conn, Nick Gould, and Ph Toint. A globally convergent lagrangian barrier algorithm for optimization with general inequality constraints and simple bounds. *Mathematics of Computation of the American Mathematical Society*, 66(217):261–288, 1997.

[61] Graziano Chesi, Andrea Garulli, Alberto Tesi, and Antonio Vicino. Solving quadratic distance problems: an lmi-based approach. *Automatic Control, IEEE Transactions on*, 48(2):200–212, 2003.

[62] Gurobi. Inc. gurobi optimizer reference manual, 2012, 2014.

[63] Andrew Makhorin. Glpk (gnu linear programming kit), 2008.

[64] IBM ILOG CPLEX. V12. 1: User's manual for cplex. *International Business Machines Corporation*, 46(53):157, 2009.

[65] Christophe H Schilling, Jeremy S Edwards, David Letscher, and Bernhard Ø Palsson. Combining pathway analysis with flux balance analysis for the comprehensive study of metabolic systems. *Biotechnology and bioengineering*, 71(4):286–306, 2000.

[66] Karthik Raman and Nagasuma Chandra. Flux balance analysis of biological systems: applications and challenges. *Briefings in bioinformatics*, 10(4):435–449, 2009.

[67] Daniel R Hyduke, Nathan E Lewis, and Bernhard Ø Palsson. Analysis of omics data with genome-scale models of metabolism. *Mol. BioSyst.*, 9(2):167–174, 2013.

[68] Patrick H O'Farrell. High resolution two-dimensional electrophoresis of proteins. *Journal of biological chemistry*, 250(10):4007–4021, 1975.

[69] Harry Towbin, Theophil Staehelin, and Julian Gordon. Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. *Proceedings of the National Academy of Sciences*, 76(9):4350–4354, 1979.

[70] Knut Wagner, Tasso Miliotis, György Marko-Varga, Rainer Bischoff, and Klaus K Unger. An automated on-line multidimensional hplc system for protein and peptide mapping with integrated sample preparation. *Analytical Chemistry*, 74(4):809–820, 2002.

[71] Oscar Puig, Friederike Caspary, Guillaume Rigaut, Berthold Rutz, Emmanuelle Bouveret, Elisabeth Bragado-Nilsson, Matthias Wilm, and Bertrand Séraphin. The tandem affinity purification (tap) method: a general procedure of protein complex purification. *Methods*, 24(3):218–229, 2001.

[72] T Fröhlich and GJ Arnold. Proteome research based on modern liquid chromatography–tandem mass spectrometry: separation, identification and quantification. *Journal of neural transmission*, 113(8):973–994, 2006.

[73] Barry R Bochner, Peter Gadzinski, and Eugenia Panomitros. Phenotype microarrays for high-throughput phenotypic testing and assay of gene function. *Genome research*, 11(7):1246–1255, 2001.

[74] David J Duggan, Michael Bittner, Yidong Chen, Paul Meltzer, and Jeffrey M Trent. Expression profiling using cdna microarrays. *Nature genetics*, 21:10–14, 1999.

[75] Priti Hegde, Rong Qi, Kristie Abernathy, Cheryl Gay, Sonia Dharap, Renee Gaspard, JE Hughes, Erik Snesrud, Norman Lee, and John Quackenbush. A concise guide to cdna microarray analysis. *Biotechniques*, 29(3):548–563, 2000.

[76] Paul A Jensen, Kyla A Lutz, and Jason A Papin. Tiger: Toolbox for integrating genome-scale metabolic models, expression data, and transcriptional regulatory networks. *BMC systems biology*, 5(1):147, 2011.

[77] Ida Schomburg, Antje Chang, and Dietmar Schomburg. Brenda, enzyme data and metabolic information. *Nucleic acids research*, 30(1):47–49, 2002.

[78] Ulrike Wittig, Renate Kania, Martin Golebiewski, Maja Rey, Lei Shi, Lenneke Jong, Enkhjargal Algaa, Andreas Weidemann, Heidrun Sauer-Danzwith, Saqib Mir, et al. Sabio-rk—database for biochemical reaction kinetics. *Nucleic acids research*, 40(D1):D790–D796, 2012.

[79] Qasim K Beg, Alexei Vazquez, Jason Ernst, Marcio A de Menezes, Ziv Bar-Joseph, A-L Barabási, and Zoltán N Oltvai. Intracellular crowding defines the mode and sequence of substrate uptake by escherichia coli and constrains its metabolic activity. *Proceedings of the National Academy of Sciences*, 104(31):12663–12668, 2007.

[80] Pascale Daran-Lapujade, Sergio Rossell, Walter M van Gulik, Marijke AH Luttik, Marco JL de Groot, Monique Slijper, Albert JR Heck, Jean-Marc Daran, Johannes H de Winde, Hans V Westerhoff, et al. The fluxes through glycolytic enzymes in saccharomyces cerevisiae are predominantly regulated at posttranscriptional levels. *Proceedings of the National Academy of Sciences*, 104(40):15753–15758, 2007.

[81] Jan Schellenberger and Bernhard Ø Palsson. Use of randomized sampling for analysis of metabolic networks. *Journal of Biological Chemistry*, 284(9):5457–5461, 2009.

[82] Walter R Gilks and Pascal Wild. Adaptive rejection sampling for gibbs sampling. *Applied Statistics*, pages 337–348, 1992.

[83] Robert L. Smith. The hit-and-run sampler: A globally reaching markov chain sampler for generating arbitrary multivariate distributions. In *Proceedings of the 28th Conference on Winter Simulation*, WSC '96, pages 260–264, Washington, DC, USA, 1996. IEEE Computer Society.

[84] Alvin C Rencher. *Methods of multivariate analysis, Second Edition*, volume 492. John Wiley & Sons, 2003.

[85] Jindan Zhou and Kenneth E Rudd. Ecogene 3.0. *Nucleic acids research*, page gks1235, 2012.

[86] Jan Schellenberger, Junyoung O Park, Tom M Conrad, and Bernhard Ø Palsson. Bigg: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. *BMC bioinformatics*, 11(1):213, 2010.

[87] Wout Megchelenbrink, Martijn Huynen, and Elena Marchiori. optgp-sampler: An improved tool for uniformly sampling the solution-space of genome-scale metabolic networks. *PLoS ONE*, 9(2):e86587, Feb 2014.

[88] B Xu, M Jahic, G Blomsten, and S-O Enfors. Glucose overflow metabolism and mixed-acid fermentation in aerobic large-scale fed-batch processes with escherichia coli. *Applied microbiology and biotechnology*, 51(5):564–571, 1999.

[89] Andrew Cameron, Emilisa Frirdich, Steven Huynh, Craig T Parker, and Erin C Gaynor. Hyperosmotic stress response of campylobacter jejuni. *Journal of bacteriology*, 194(22):6116–6130, 2012.

[90] Heinrich J Huber, Heiko Dussmann, Seán M Kilbride, Markus Rehm, and Jochen HM Prehn. Glucose metabolism determines resistance of cancer cells to bioenergetic crisis after cytochrome-c release. *Molecular systems biology*, 7(1), 2011.

[91] BERTRAND Perroud and DANIEL Le Rudulier. Glycine betaine transport in escherichia coli: osmotic modulation. *Journal of Bacteriology*, 161(1):393–401, 1985.

[92] Yuichi Taniguchi, Paul J Choi, Gene-Wei Li, Huiyi Chen, Mohan Babu, Jeremy Hearn, Andrew Emili, and X Sunney Xie. Quantifying e. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329(5991):533–538, 2010.

[93] Yoshiharu Inoue, Yoshiyuki Tsujimoto, and Akira Kimura. Expression of the glyoxalase i gene of saccharomyces cerevisiae is regulated by high osmolarity glycerol mitogen-activated protein kinase pathway in osmotic stress response. *Journal of Biological Chemistry*, 273(5):2977–2983, 1998.

[94] Orna Carmel-Harel and Gisela Storz. Roles of the glutathione-and thioredoxin-dependent reduction systems in the escherichia coli and saccharomyces cerevisiae responses to oxidative stress. *Annual Reviews in Microbiology*, 54(1):439–461, 2000.

[95] DO Bayles and BJ Wilkinson. Osmoprotectants and cryoprotectants for listeria monocytogenes. *Letters in applied microbiology*, 30(1):23–27, 2000.