NATIONAL TECHNICAL UNIVERSITY OF ATHENS
SCHOOL OF CHEMICAL ENGINEERING
DEPARTMENT OF PROCESS ANALYSIS AND PLANT DESIGN

On the integration of process engineering with metabolomics for the production of muconic acid: the case for *Saccharomyces cerevisiae*

Thesis submitted for the degree of

Masters in Chemical Engineering

By

Stefanos Xenios

Supervisors: Professor Antonis Kokossis (NTUA)

Professor Vassily Hatzimanikatis (EPFL)

Athens, February 2021

Ευχαριστίες

Η εκπόνηση της παρούσας διπλωματικής εργασίας έγινε σε συνεργασία με το εργαστήριο του καθηγητή Βασίλη Χατζημανικάτη στο EPFL, τον οποίο ευχαριστώ θερμά για την ευκαρία που μου έδωσε να ασχοληθώ με το κομμάτι της μεταβολικης μηχανικής. Θα ήθελα να ευχαριστήσω τον διδάκτορα του EPFL Daniel Weilandt για την υποστήριξη και την συνεχή παρακολούθηση και βοήθεια στο κομμάτι της ανάπτυξης των κινητικών μοντέλων. Οφείλω ένα μεγάλο ευχαριστώ στον επιβλέποντα καθηγητή μου Αντώνη Κοκόση για την καθοδήγηση και την βοήθεια του στο κομμάτι της μηχανικής μάθησης.

Ακόμα θα ήθελα να ευχαριστήσω την οικογένεια μου που βρίσκεται δίπλα μου και με στηρίζει σε οποιοδήποτε εγχείρημα μου. Χωρίς εσάς δεν θα ήταν τίποτα δυνατό.

Abstract

Muconic acid is a high value product which has gathered interest in applications in the manufacture of new resins, bio plastics, food additives, agrochemicals and pharmaceuticals. Lots of efforts have been made for an economically viable biotechnological strategy for muconic acid production but as of yet have been fruitless. Directed evolution and DBTL cycles hold important promises for the development of future catalysts with high efficiency and productivity. However, process engineering is typically disjointed from these cycles and more often than not the mismatch of kinetics presents a major challenge and a bottleneck in the scaling up of novel bioprocesses.

The dissertation addresses the integration of metabolomics and experimental data using the optimization and risk analysis of complex living entities (ORACLE) platform combined with clustering and advanced analytics. The methodology consists of six steps. In the first step, the stoichiometry of the system is defined through biochemical data and experimental data are integrated into the model to further constrain it. In the second step, steady state fluxes and metabolite concentrations are calculated based on metabolomics analysis. In the third step, through stoichiometric analysis conserved moieties are identified and the dependent metabolites are separated from the independent. In the fourth step, kinetic parameters for every reaction are sampled to fit in with the steady state fluxes based on mechanistic kinetics expressions. In the fifth step, consistency checks and pruning consider the stability of the system and the consistency with experimental data. In the fifth step, the flux control coefficients for the desired metabolite flux are calculated based on the well-established metabolic control analysis (MCA) framework. In the sixth step, clustering and advanced statistical analysis on the control coefficient population is performed to determine the impact of key enzymes on the desired flux.

In this project, large-scale mechanistic kinetic models for a muconic acid producing S.cerevisiae strain were developed using the aforementioned ORACLE platform. The yeast8 genome scale model [1] was used and experimental data from this paper [2] were integrated into the model. Three heterologous reactions (PaAroZ, KpAroY, CaCatA) were added to the GEM for the muconic acid production pathway via shikimate pathway branching. The reduced genome scale model for S.cerevisiae used in this project consisted of 306 reactions and 300 metabolites. A total of 23500 of potential kinetic models were generated out of which 372(1.58%) agreed to the experimental data thus passing the pruning step. Lastly 29(0.12%) models out of the 372 passed the consistency check and showed stability in random perturbations performed on them. Those 29 models were used to indicate key enzymes that affect muconic acid flux and possible bottlenecks. Enzyme perturbations were performed to further quantify the influence of various enzymes on muconic flux. A big number of enzymes seem to have a significant impact on muconic acid production, excluding those that express the heterologous reactions of the muconic acid pathway, such as glucose-6-phosphate isomerase (PGI), transketolase (TKT2) and enolase (ENO).

This study aims to offer metabolic engineering strategies for a muconic acid production yeast strain while taking into consideration stoichiometry, thermodynamics and kinetics.

Keywords: *S.cerevisiae*; muconic acid; genome scale model; metabolomics; advanced data analytics; metabolic engineering; ORACLE; metabolic control analysis; large-scale kinetics; industrial biotechnology

Εκτεταμένη περίληψη

Η παρούσα διπλωματικη εργασία πραγματοποιήθηκε στα πλαίσια της ακαδημαικής συνεργασίας μεταξύ της σχολής Χημικών Μηχανικών Ε.Μ.Π και του πανεπιστημίου EPFL. Η εργασία είχε ως στόχο την ανάπτυξη μεγέλου μεγέθους κινητικών μοντέλων με υπολογιστικές μεθόδους για στελέχη του μύκητα S.cerevisiae με την ικανότητα να παράγουν μυκονικό οξύ.

Η χρήση γενετικά τροποποιημένων μικροοργανισμών για την παραγωγή βιοκαυσίμων και χημικών αποτελεί μια υποσχόμενη βιώσιμη εναλλακτική. Η δημιουργία κατάλληλων στελεχών ικανών να παράγουν χρήσιμα χημικά συχνά προϋποθέτει την επιβολή αλλαγών στον κυτταρικό μεταβολισμό. Επειδή οι αλλαγές αυτές είναι μη προφανείς η σχεδιαστική διαδικασία συχνά υποβοηθάται από τη χρήση Μεταβολικών μοντέλων Γονιδιακής Κλίμακας (ΜΓΚ) που περιέχουν όλη τη διαθέσιμη πληροφορία σχετικά με τις μεταβολικές δυνατότητες ενός οργανισμού. Τα ΜΓΚ αποτελούν στοιχειομετρικές αναπαραστάσεις του συνόλου του μεταβολικού δικτύου υπό τη μορφή συστήματος γραμμικών εξισώσεων και περιορισμών. Τα ΜΓΚ αποτελούν μία απλοποίηση του πολύπλοκου συστήματος ενός μικρορ̣γανισμού καθώς χρησιμοποιούν την παραδοχή της μόνιμης κατάστασης. Η παραδοχή αυτή καθιστά δυνατή την γρήγορη εξαγωγή συμπερασμάτων και μεταβολικών στρατηγικών ως προς τι γενετικές τροποποιήσεις που θα προσδώσουν στον μικροοργανισμό κάποιο επιθυμητό χαρακτηριστικό. Ωστόσο αυτή η μέθοδος δεν λαμβάνει υπόψιν την δυναμική του συστήματος, δηλαδή την κινητική και αδυνατεί να εντοπίσει πιθάνα τροχοπέδια του μεταβολικού συστήματος και να υποδείξει με μεγαλύτερα σιγουριά ένζυμα στόχους που επηρεάζουν την παραγωγή του επιθυμητού προιόντος.

Η ανάπτυξη κινητικών μοντέλων ενός συστήματος τέτοιας πολυπλοκότητας όπως ένας μικροοργανισμός είναι μια αρκετά δύσκολη πρόκληση. Η κινητικη των ενζύμων εργαστηριακά καθορίζεται με in vitro πειράματα που ο ενζυμικος καθαρισμός είναι υποχρεωτικός. Ακόμα και για τον πιο μελετημένο μικροοργανισμό S.cerevisiae δεν έχουν προσδιοριστεί οι κινητικές παράμετροι της πλειονότητας των ενζύμων. Αν και γίνονται πειράματα προσδιορισμού κινητικών ιδιοτήτων των ενζύμων ποτέ δεν θα επαρκέσουν για να περιγράψουν όλη την δυναμική συμπεριφορά του πολύπλοκου μεταβολικού συστήματος αποτελείτο απο εκατονταδες ένζυμα.

Μία άλλη στρατηγική είναι να προσπαθήσουμε να υποθέσυμε κινητικές εκφράσεις για τις ενζυματικές αντιδράσεις και να προβλέψουμε τις κινητικές παραμέτρους κάθε αντίδρασης. Η πολυπλοκότητα των ενζυμικών αντιδράσεων ωστόσο, καθιστά σχεδόν αδύνατο τον χαρακτηρισμό όλων των κινητικών παραμέτρων. Επομένως, συνήθως χρησιμοποιούνται παραδοχές και απλουστεύσεις στις εκφράσεις των ενζυμικών κινητικών, είτε κινητικές Michaelis-Menten είτε κινητικές δράσης μαζών. Τις περισσότερες φορές τα κινητικά μοντέλα υποθέτουν τις κινητικές παραμέτρους δίχως πειραματικά δεδομένα ή δεν λαμβάνουν υπόψιν ρυθμούς αντίδρασης μόνιμης κατάστασης ή δεδομένα συγκεντρώσεων, ούτε εξασφαλίζουν θερμοδυναμικούς περιορισμούς.

Στην παρούσα εργασία χρησιμοποιήθηκε η μεθοδολογία ORACLE(Optimization and Assessment of Complex Living Entities) για την ανάπτυξη κινητικών μοντέλων για γενετικά τροποποιημένα στελέχη μαγιάς που παράγουν μυκονικό οξύ. Στο πρώτο στάδιο ενσωματώνονται στο επιλεγμένο Μεταβολικο μοντέλο Γονιδιακής Κλίμακας πειραματικά δεδομένα αν υπάρχουν, όπως και οι ετερόλογες ενζυμικές αντιδράσεις. Τα ΜΓΚ αποτελούν στοιχειομετρικές αναπαραστάσεις του συνόλου του μεταβολικού δικτύου υπό τη μορφή συστήματος γραμμικών εξισώσεων και

περιορισμών. Στο δεύτερο στάδιο δειγματοληπτείται ο επιτρεπτός χώρος λύσεων του γραμμικού συστήματος και δειγματοληπτούνται επιτρεπτές συγκεντρώσεις μεταβολιτών. Στο τρίτο βήμα, υπολογίζονται οι κινητικές παράμετροι των αντιδράσεων έτσι ώστε να επαληθεύουν τις λύσεις της μόνιμης κατάστασης. Στο τέταρτο βήμα, τα κινητικά μοντέλα που έχουν παραχθεί υποβάλονται σε τεστ που ελέγχουν την σταθερότητα τους σε τυχαίες μεταβολές της αρχικής συγκέντρωσης ισορροπίας καθώς και την ταχύτητα απόκρισης. Τα κινητικά μοντέλα που εμφανίζουν υψηλή σταθερότητα χρησιμοποιούνται για τον υπολογισμό συντελεστών ρύθμισης μεταβολικής ροής και συντελεστών ρύθμισης μεταβολικών συγκεντρώσεων σύμφωνα με την μεθοδολογία MCA(Metabolic Control Analysis). Στο τελευταίο βήμα, εξάγονται στατιστικά δεδομένα για τους συντελεστές ρύθμισης και προτείνονται μεταβολικές στρατηγικές αύξησης ροής του επιθυμητού προιόντος.

Με την μεθοδλογία αυτή καταφέραμε για το σύστημα μας καταφέραμε να αναπτύξουμε 70 σταθερά κινητικά μοντέλα. Επίσης εντοπίστηκαν ένζυμα στόχοι και προτάθηκαν μεταβολικές στρατηγικές που επηρεάζουν σε υψηλό βαθμό την παραγωγή μυκονικού οξέος και παρατηρήθηκε εώς και 12πλασιασμός της επιθυμητής μεταβολικής ροής. Οι μεταβολικές στρατηγικές μπορούν να μεταφραστούν σε εργαστηριακές πρακτικές γενετικής τροποποίησης όπως προσθήκη, διαγραφή, υπερέκφραση ή υποέκφραση γονιδίων και να θέσουν στόχους για την τροποποίηση ενζύμων.

Table of content

List of Figures

List of Tables

Chapter 1. Introduction

Biotechnology has existed almost since the dawn of humanization. Products such as cheese, bread, wine and beer come from the manipulation of living organisms. As humans began to understand better biology and the possibilities of bioprocesses and with the sense of necessity born from both Worlds Wars, the first biorefineries were constructed. The acetone-ethanol-butanol fermentation process, which is still in use, was developed during WW1 and WW2 was signaled the industrial scale production of penicillin[1].

In the early 1990s advancement in the field of genetic engineering led to success stories especially in the pharmaceutical field (recombinant proteins, antibodies). Later, mathematical modelling, genome scale models, easier genetic modification technologies, bioinformatics kickstarted the field of metabolic engineering- suggesting was of producing chemicals exceeding the spectrum of the food industry through biotechnological routes. The emerging biotechnological field has a net worth of 300 billion USD and is said to duplicate by 2025[2].

Biotechnology could prove to be a solution to a plethora of modern-day challenges such as climate change, pollution, depletion of natural resources and increasing food demands. Biotechnology is an invaluable that will be able to transform renewable feedstock to desired chemicals thus decreasing the need for fossil fuels and subsequently alleviating the impact on climate. Genetically modified plants would prove to be highly resistant and exhibit high yields helping in the rising food demands. Biotech drugs, vaccines and diagnostics help rise the quality of life[3,4].

Bioprocesses and biorefineries where biomass feedstock are transformed to biofuels, platform and special chemicals and novel products, can be seen as the key concepts that will lead to a bio-based economy. Fossil fuels and petrochemicals still dominate today's economy as the cost of production is lower than their bio counterparts. With the ever-increasing fuel and plastic demand and the stagnation of old oil wells, petroleum prices will soon rise as we will have to turn to much more inaccessible oil wells and more unrefined oil. This will result in higher extraction costs and higher separation and purification costs. On the other hand, bio processes remain an environment friendly alternative, as microbial fermentations demand milder conditions than the equivalent catalytic transformations. However, these processes are seldom competitive to their petrochemical counterparts due to low productivity, yields and high separation costs. The development of industrial strains with selected characteristics that can support the commercialization of a biorefinery application is being conducted with iterative cycles where metabolic interventions are systematically identified and applied to the host organism[5-7]. As the metabolic systems are very complex, interventions are not so obvious and are computer assisted and typically Mixed Integer Linear Programming (MILP) algorithms are utilized to build the interventions[8-10].

However, most strain design approaches make the assumption of the system being in steady state and fail to take into consideration the kinetic aspect. Although such techniques provide useful insights and are computationally cost efficient, sometimes they fail to predict potential bottlenecks or rate limiting reactions inside the metabolic network. The current thesis aims to develop large scale metabolic kinetic models for a muconic acid producing yeast using the Optimization and Risk Assessment of Complex Living Entities (ORACLE) workflow. The generated populations of kinetic models will be used as an input for the well-established MCA framework to

identify enzymes closely affiliated to muconic acid flux and perform enzyme perturbations to quantify this effect. This way we will be able to develop metabolic strategies while taking into consideration stoichiometry, thermodynamics, kinetics and their interplay.

Chapter 2. Background and state of the art

2.1 Cellular Metabolism

Metabolism refers to all chemical reactions that take place in an organism. These reactions are necessary for the production of biomolecules and biomass (nucleic acids, lipids, proteins and carbohydrates), the extraction of energy from nutrients for the various functions of the cell. Metabolism is often divided into catabolism and anabolism.

Catabolism is the set of reactions that break down large molecules (such as polysaccharides, lipids nucleic acids and proteins) into smaller units (such as monosaccharides, fatty acids, nucleotides and amino acids respectively). These smaller units are either oxidized to release energy necessary for the maintenance and growth of cells or used in anabolic reactions. Some examples of catabolic pathways are glycolysis, the citric acid cycle, the breakdown of fat or muscle protein and many more. Glycolysis for example is the metabolic pathway that converts glucose into pyruvic acid. During this process free energy is produced and while a large part escapes as heat the other is used to form high-energy molecules ATP (adenosine triphosphate) and NADH (reduced nicotinamide adenine dinucleotide).

Anabolism, or also known as biosynthesis, on the other hand is the set of metabolic pathways that construct large molecules from smaller units. This process requires energy and is called endergonic whereas catabolism is exergonic. Some examples of anabolism are the amino acid biosynthesis, gluconeogenesis etc.

The sum of the aforementioned reactions, catalyzed by enzymes, consist the metabolism of the cell. These reactions form interconnected metabolic pathways that shape a dynamical circuitry referred as metabolic network. The pathways are further categorized in subnetworks, responsible for a specific cellular function. Organisms have many structural differences between them but they may share several core subnetworks.

2.2 Metabolic Engineering

Metabolic engineering is the intentional modification and manipulation of cellular metabolism for the production of desired chemicals[11]. Recombinant DNA techniques are used to either insert new pathways in microorganisms in order to produce novel metabolites or produce heterologous peptides (e.g., human insulin) or improve new or existing processes.

Metabolic engineering is necessary for creating efficient cell factories for the development of sustainable processes for the production of chemicals, fuels and materials. Microorganisms through their complex metabolic network are able to convert simple feedstock such as glucose, lignin etc. into desired products. However small yields, high separation costs, lack of kinetics, difficulties in the scaling up impose a bottleneck. Systems metabolic engineering can offer better predictions by combining the omics data (genomes, transcriptomes, proteomes, metabolomes, fluxomes) and computational techniques used in systems biology allowing a better understanding of the cellular processes and engineering capabilities[12].

2.2.1 Muconic acid

Muconic acid (MA) is a high value product which has gathered interest in applications in the manufacture of new resins, bio-plastics, food additives, agrochemicals and pharmaceuticals[13]. MA is an unsaturated dicarboxylic acid, also known as 2,4 hexanedienoic acid, which due to the double bond and dicarboxylic groups can be polymerized. MA can be found in three isomer forms cic,cis-MA, cis,trans-MA, trans,trans-MA. MA can be chemically processed to produce bulk chemicals such as adipic acid, terephthalic acid and trimelitic acid which are widely used in the nylon and thermoplastic industry[13].

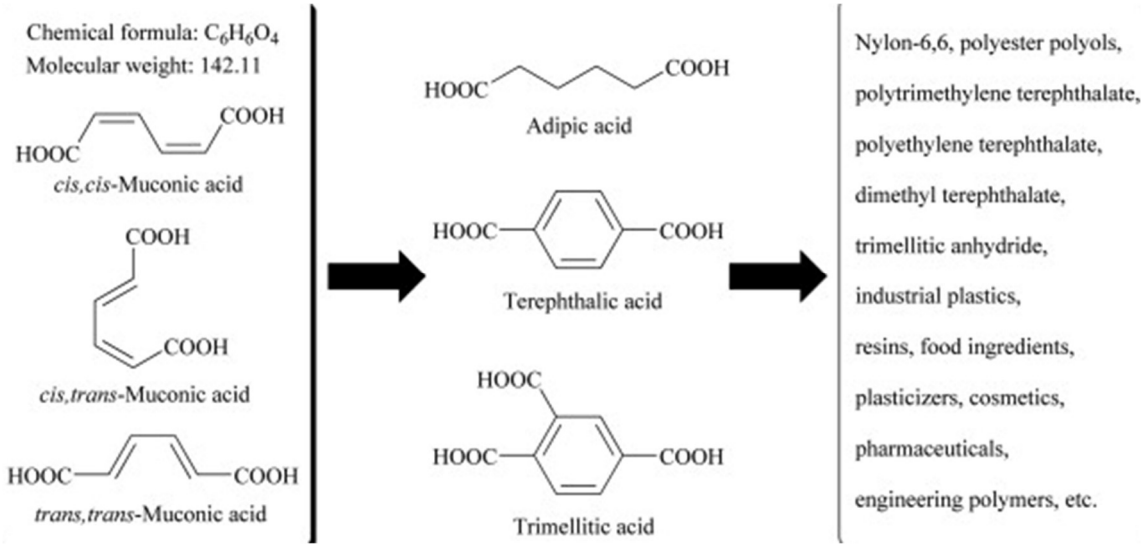

Figure 1:Industrial products of MA and further applications[13]

Traditionally MA is produced via chemical processing of non-renewable petroleum feedstock with acids and strong metal catalysts[14]. Toxic intermediates, corrosive catalysts, concern for the environment and the use of non-renewable materials have led to the search of alternative paths. Recently, progress has been made in the biochemical procedures for MA production.

2.2.2 Different Metabolic strategies for muconic acid production

The ortho cleavage of catechol

Catechol is a precursor molecule to cis,cis-MA and can be produced by microorganisms capable of biodegrading aromatic compounds. Especially, bacteria oxidize aromatic compounds such as benzoate, toluene, benzene, phenol, aniline, anthranilate to catechol. Benzoate has mainly been used as feedstock due to its low price and the capability of some microorganisms in the genus Pseudomonas, Arthrorbacter, Corynebacterium, Brevibacterium, Microbracterium and Sphingobacterium to metabolize to MA via the β-ketoadipate pathway[15].

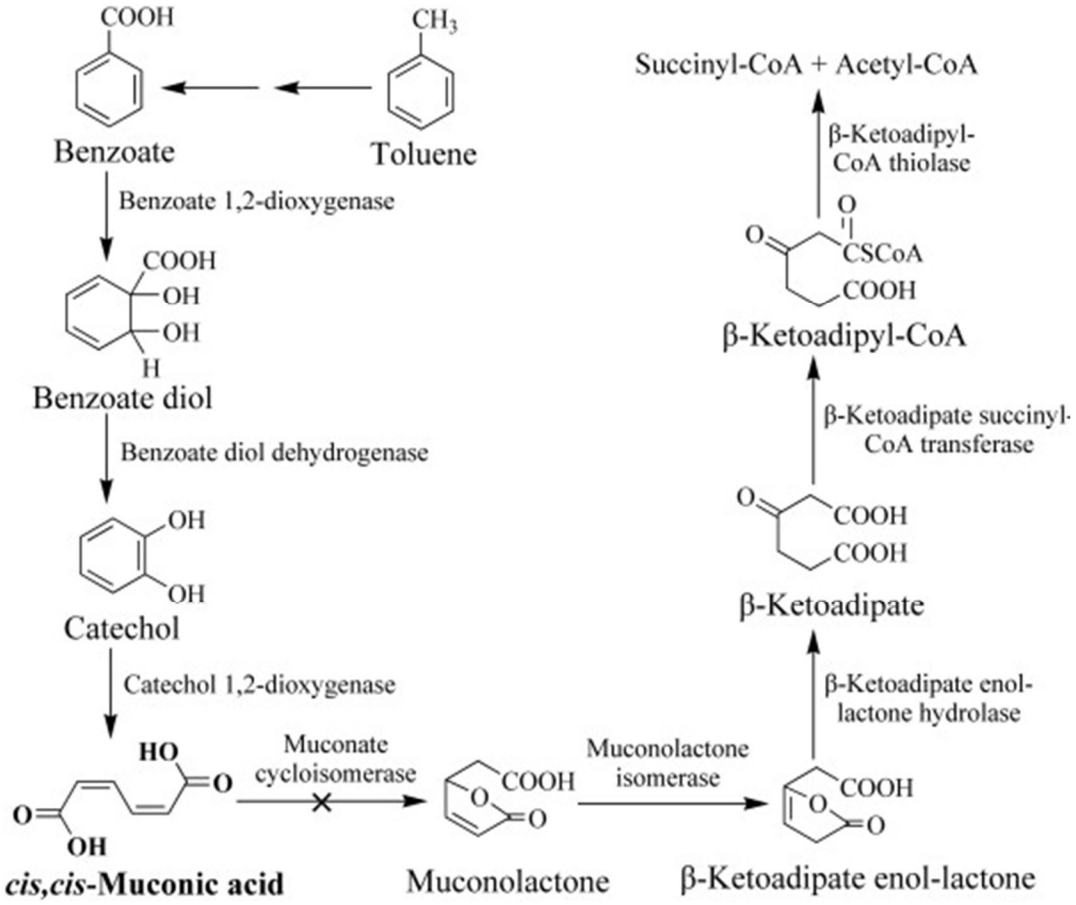

Figure 2:The β-ketoadipate pathway[15]

In order to achieve high MA yields this pathway requires the following properties for the mutants;

- Metabolizing aromatic compounds through the β-ketoadipate pathway and having the CatA enzyme responsible for the cleavage of the aromatic ring of catechol
- Missing functional muconate cycloisomerase which would enable the accumulation of MA and stop the flux to muconolactate
- Being resistant to aromatic products and substrates
- Excreting MA to the medium in order for it to be easily separated
- Having strong CatA activity[16]

Many efforts have been made with the above biotechnological strategy and high yields, concentrations, productivities have been achieved[17-18]. However, this method is restrictive due to its dependency on the petrochemical industry. An attractive alternative would be the use of a lignocellulosic feedstock due to its high concentration of aromatic compounds[19-20] but the high cost of separation and small yields present a bottleneck.

MA production by branching shikimate synthesis via 3-dehydroshikimate

Glucose can be converted to MA through an artificial pathway based on the shikimic pathway. The shikimic pathways begins with the condensation of phosphoenolpyruvate (PEP) and erythrose-4-phosphate to form 3-deoxy-D-arabino-heptulosonate-7-phosphate (DAHP). DAHP is

then converted to 3-dehydroquinate (DHQ) and afterwards to dehydroshikimate (DHS). DHS is the precursor molecule to shikimate which is the basic aromatic compound for the production of aromatic amino acids such as tryptophan, phenylalanine, tyrosine. An heterologous pathway is expressed by the genes aroZ (encodes 3-dehydroshikimate dehydratase which catalyzes the reaction of DHS to protocatechuate (PCA)), aroY(encodes protocatechuate decarboxylase which catalyzes the reaction of PCA to catechol) and catA(encodes catechol dioxygenase which catalyzes the reaction of catechol to MA).

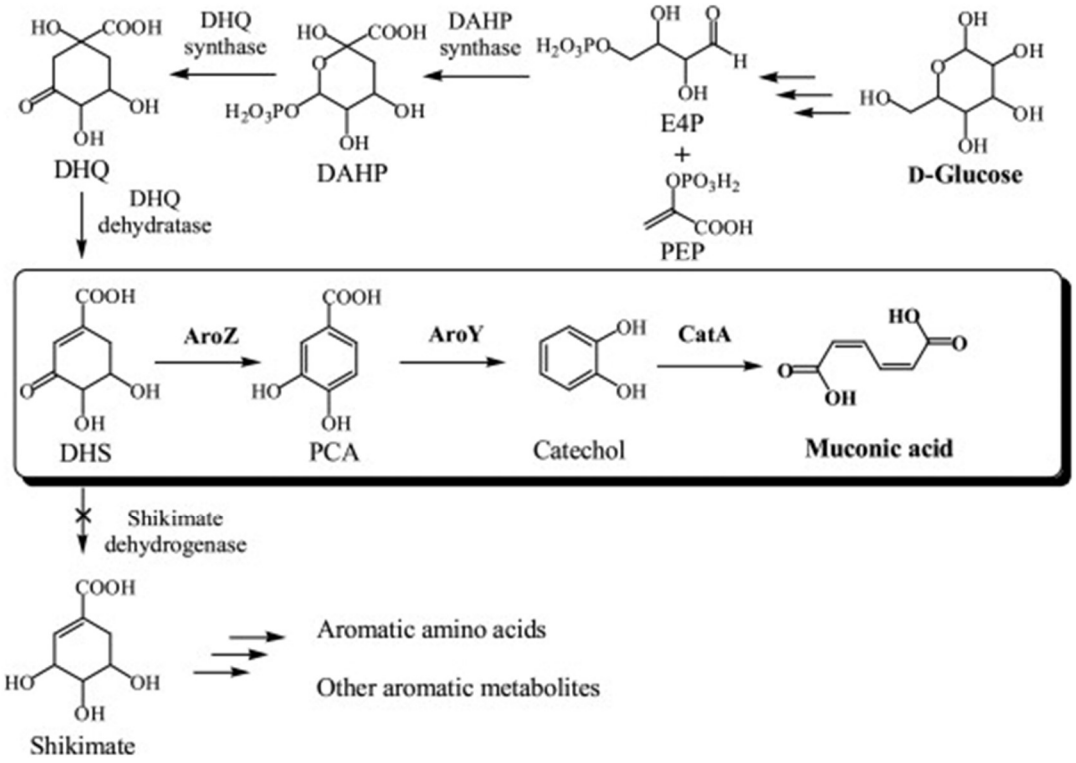

Figure 3:The heterologous pathway to MA production by branching of the shikimic pathway[13]

Many efforts have been made using a shikimate dehydrogenase deficient E.coli. It is worth mentioning. E.coli is a great candidate for high yield muconic acid production. An engineered E.coli strain with inactivated shikimate dehydrogenase, overexpression of DHQ synthase and transketolase to increase the flux through the shikimic pathway, and the heterologous enzymes AroZ and AroY and CatA has produced 59g/L(30% yield mol/mol) in a fed-batch culture, the highest reported MA yield with glucose as a substrate[21].

MA production by branching tryptophan pathway via anthranilate

A novel alternative to the branching of the shikimic pathway is the branching of the tryptophan pathway. More specifically, an intermediate metabolite of the shikimic pathway is chorismate which is a precursor molecule to the three aromatic amino acids tryptophan, phenylalanine, tyrosine. The reaction catalyzed by anthranilate synthase converts chorismate to anthranilate. A novel pathway is grafted by heterologously expressing the reaction of anthranilate to catechol encoded by the gene ADO. Catechol then gets converted to MA through the CatA enzyme activity.

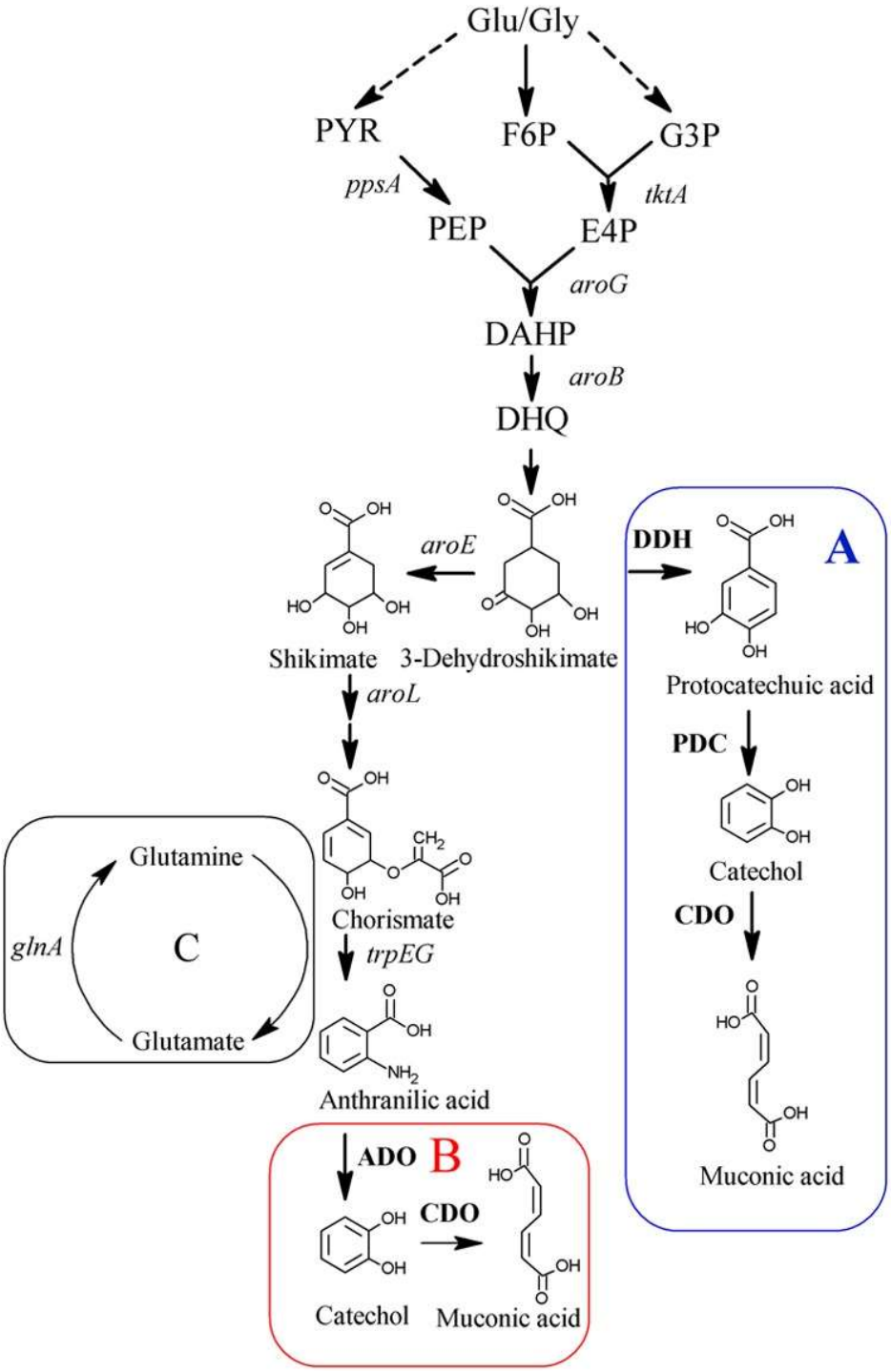

Figure 4:Pathway A illustrates muconic acid production via shunting the shikimic pathway and pathway B illustrates muconic acid production via shunting antranilate pathway[22]

An engineered E.Coli strain following the above pathway produced a maximum concentration 390mg/L of MA in shake flasks[22]. This heterologous pathway is shorter by one reaction from the shikimate alternative but additional PEP is required in contrast[23-24] and small yields have as of now been reported.

2.2.3 Yeast as a cell factory

Since S.cerevsiae has been broadly used for beer and wine fermentations, its selection as industrial ethanol producer is not surprising. The ethanol fermentation, nowadays, consists a robust, well-studied industrial application, making yeast one of the most preferred host organisms for the production of diverse fuels and chemicals. Moreover, holding a GRAS status by FDA, makes yeast suitable for the production of food-grade products[25]. Over the years, yeast metabolic capabilities have been exploited for the production of various products such as pharmaceuticals (artemisinic acid, human albumin etc.), fuels (alcohols, alkanes, etc.) and platform chemicals (succinic acid, coumaric acid, etc.) and specialties (santalene, valencene, etc.).

Many microorganisms are capable of producing muconic acid naturally using aromatic compounds such as lignin as feedstock. However, the production levels and properties fall short due to toxicity, different feedstock compositions, difficult and costly separations. Although, microorganisms such as *K.putida* have shown very high yields with pure aromatic compounds as feedstocks they fail to provide a viable solution due to the high value of the materials used. For that reason, interest is being shown to hosts that can be used to express heterologous pathways that would lead to a cost-efficient biotechnological strategy using cheap feedstocks.

Yeast, although falling short to its bacteria adversaries *K.putida* and *E.coli* who can produce and metabolize aromatic compounds due to their physiology, is the perfect host to express heterologous pathways. Moreover, the accumulation of yeast fermentation data and knowledge reduce the uncertainty in the scaling up process. S.cerevisiae exhibits pH-tolerance and has proven robust in prior applications. Yeast from a biological scope is the perfect candidate for gene editing as:

- Its genome is completely sequenced, its metabolic pathways are known as well as its proteins[26]
- Genetic manipulation on yeast is quite easy as DNA insertion and integration in a yeast cell doesn't require transport proteins or CRISPR
- Yeast cells don't contain a cell membrane thus making it easier to insert plasmids.

2.3 Constraint based modelling

The generation of large amounts of information regarding biological systems and processes due to emerging high throughput technology has also pushed forward the systematic and mathematical analysis of these systems[27]. The mathematical modelling of metabolism is an invaluable tool in predicting and evaluating the cellular reaction under genetic changes, concentration perturbations and different background environments. In general, all the existing approaches boil down to roughly two categories: the kinetics modelling and the stoichiometric modelling. The kinetics approach is based in assigning mechanistic kinetic expressions such as Michaelis Menten or Hill equation or mass action kinetic expressions to describe the reaction rates. The conservation of mass for every metabolite depicts an ordinary differential equation. The total of these equations presents a system of ordinary differential equations which can be solved numerically. The solution of such a system is a time dependent metabolite concentration and reaction flux profile. However, lack of kinetics experimental data or uncertainty in the prediction of kinetics parameters present an obstacle. The stoichiometric approach primarily

relies in the assumption of steady state resulting in a system of linear equations that describe the metabolites mass concentration.

2.3.1 Genome scale metabolic reconstructions (GEMS)

Stoichiometric models have been used to study the physiology of organisms for almost 30 years. The advance in technology as well as the better understanding of genomes has led to the accumulation of information and the creation of Genome Scale Metabolic Reconstructions[28]. GEMS can be seen as a map showing all the reactions occurring in the cell while also linking the gene encoding the enzyme for each reaction. This gene to protein to reaction association (GPR) offers a pretty good image of the biochemistry inside the cell. These GEMS can be used for five major ends: (1) conceptualization of high-throughput data, (2) assisting in metabolic engineering, (3) directing hypothesis-driven discovery, (4) evaluation of multi-species interactions, (5) network property analysis[29]. For metabolic engineering applications GEMS can offer metabolic strategies for maximizing a desired product flux, identification and drug design as well as reactions of cellular phenotypes under different environment conditions and gene knockouts.

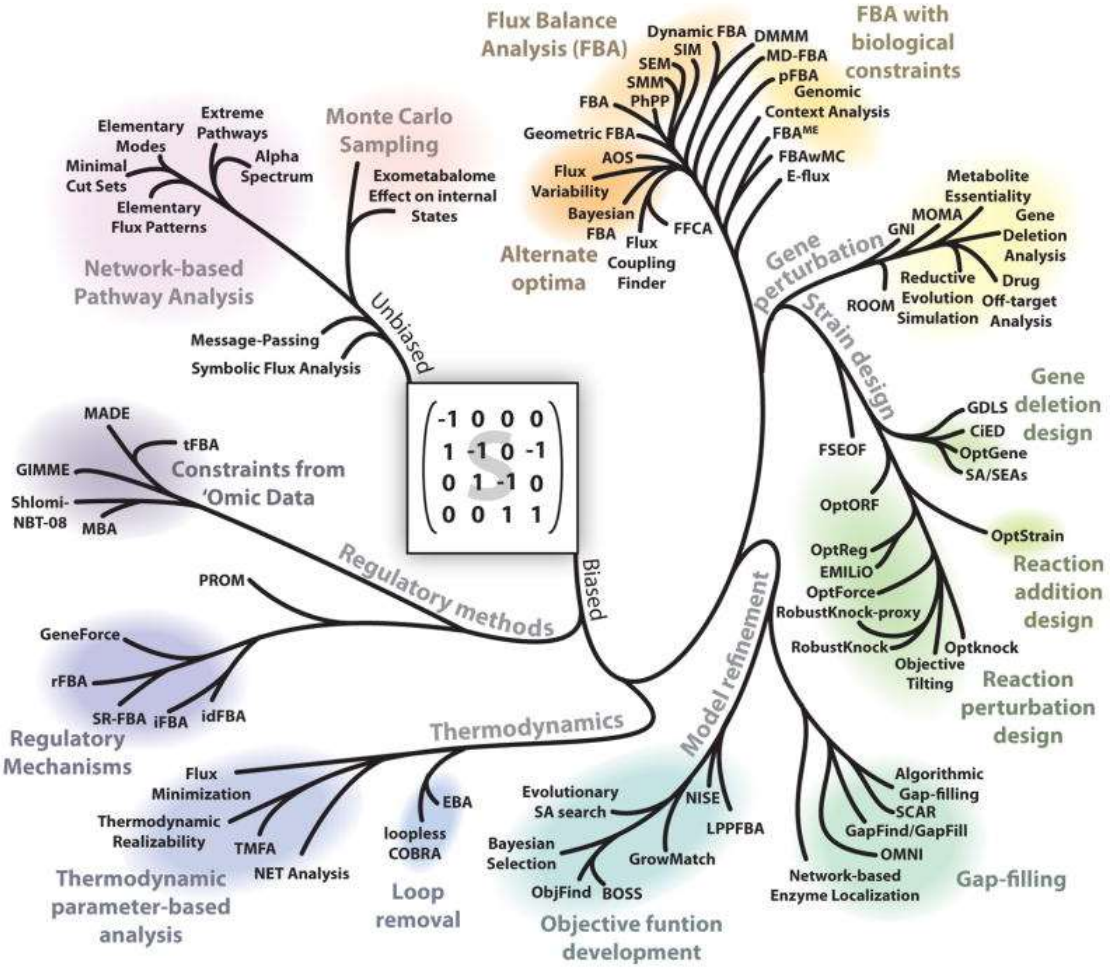

Figure 5:Phylogenetic representation of the alternative constraint-based methods applied to GEMs[30]

2.3.2 Flux Balance Analysis (FBA)

Flux Balance Analysis is a key method for constraint-based modelling from which take birth a plethora of other analysis methods. At its core, FBA is a mathematical approach to analyze the flow of metabolites through a metabolic pathway[31]. The metabolic network stoichiometric information is encoded in the stoichiometric matrix S. Each row of the matrix S represents a metabolite and its column a reaction. The elements of the matrix are the stoichiometric coefficients of every metabolite for every reaction. FBA imposes two constraints on the system. The first one is that the system is in pseudo steady state, meaning that there exists none time dependency. This is mathematically described by the equation:

$$S \cdot v = 0$$

Where S is the stoichiometric matrix and v is the flux vector for every reaction. This simple equation makes sure that every mass concentration equation for every metabolite is satisfied. The second type of imposed constraints are the lower and upper bounds for metabolite concentration, typically including laboratory flux measurements (metabolite uptake or secretion rates).

These two types of constraints define am allowable solution space. The network is capable to acquire any flux distribution lying inside the solution space. Typically, the aim of FBA is to find a flux distribution inside the allowable solution space that maximizes or minimizes a specified objective function. A common aim when handling GEMS is growth prediction, so a column is added most of the times to account for the biomass producing reaction. The objective function in this case is the maximization of biomass production.

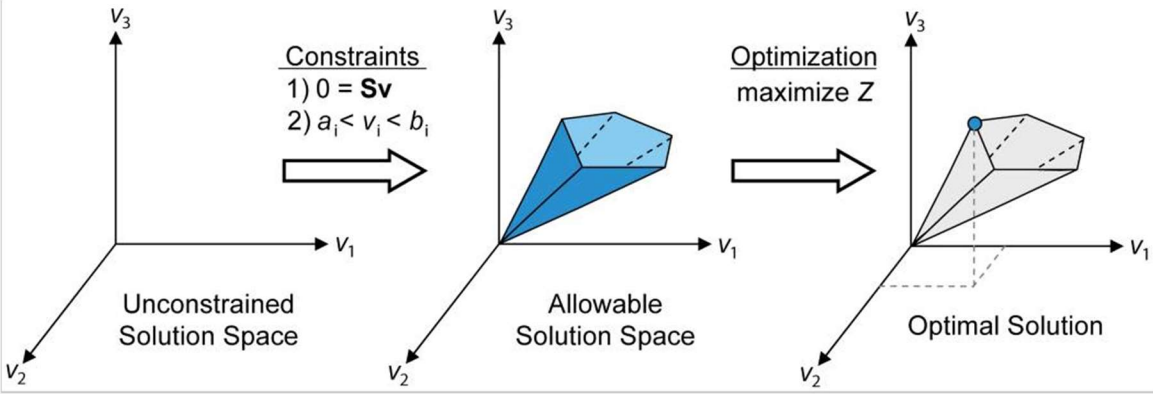

Figure 6:: A conceptual basis representation of constraint-based modeling. When left unconstrained any flux profile is possible. When mass balance constraints (S matrix) and lower and upper bounds of metabolites are defined, then there exists an allowable solution space. The network may acquire any flux distribution within this defined space, while points outside it are denied by the constraints. Through optimization of specific objective functions, FBA identifies a point in the allowable solution space that satisfies the objective function.

2.3.3 Thermodynamic Flux Analysis (TFA)

While FBA solutions offer a good insight in metabolic fluxes more often than ever the solutions are non-unique and sometimes violate thermodynamic laws. To further constrain the solution space and obtain thermodynamically feasible flux profiles extra constraints are added in order to

couple the reaction directionalities to thermodynamic constraints[32-35]. In this approach metabolite concentrations and Gibbs energy of reaction are added to the model.

Table 1.FBA and TFA constraints[35]

FBA constraints	Mass balance	$S \cdot v = 0$
	Flux capacity	$\bar{v} \leq \mathrm{v} \leq \underline{v}$
TFA constraints	Gibbs energy of reaction	$\Delta_r G_i = \Delta_{r,tpr} G_i' + \sum_{j=1}^{m} n_{i,j} m_j$
	Chemical potential	$\mu_j = \Delta_f G_j'^0 + \Delta_{f,err} G_j'^0 + RTlnx_j$
	Thermodynamic feasibility	$\Delta_r G_i' - K + K \cdot z_i < 0$
	Coupling constraint	$v_i - K \cdot z_i < 0$

Where:

ΔrGi ' is the transformed Gibbs free energy of the reaction i

μj are the chemical potentials of the reactants j

Δ$r,tptGi$ ' is the Gibbs free energy of transport (accounted when the reaction is transport of a compound from one compartment to another)

ΔfGj '0 is the standard transformed Gibbs free energy of formation of the compounds

Δ$f,errGj$ '0 is the estimated error in the energy of formation

R is the universal gas constant

T is the temperature (here assumed 298 K)

xj is the molar fraction of the compound j

K is a large (Big-M, K > max ΔrGi ') value

and z is a binary decision variable

This formulation further requires that net fluxes are non-negative. To achieve this, each reaction is separated in two: a net forward and a net backward while the net fluxes are associated such that:

$$v_{net} = v_{forward} - v_{backward}$$

By doing this it is ensured that the solution space is a convex one which is necessary during the sampling step. From the constraints it is also ensured that either the backward reaction is active or the forward or none at all.

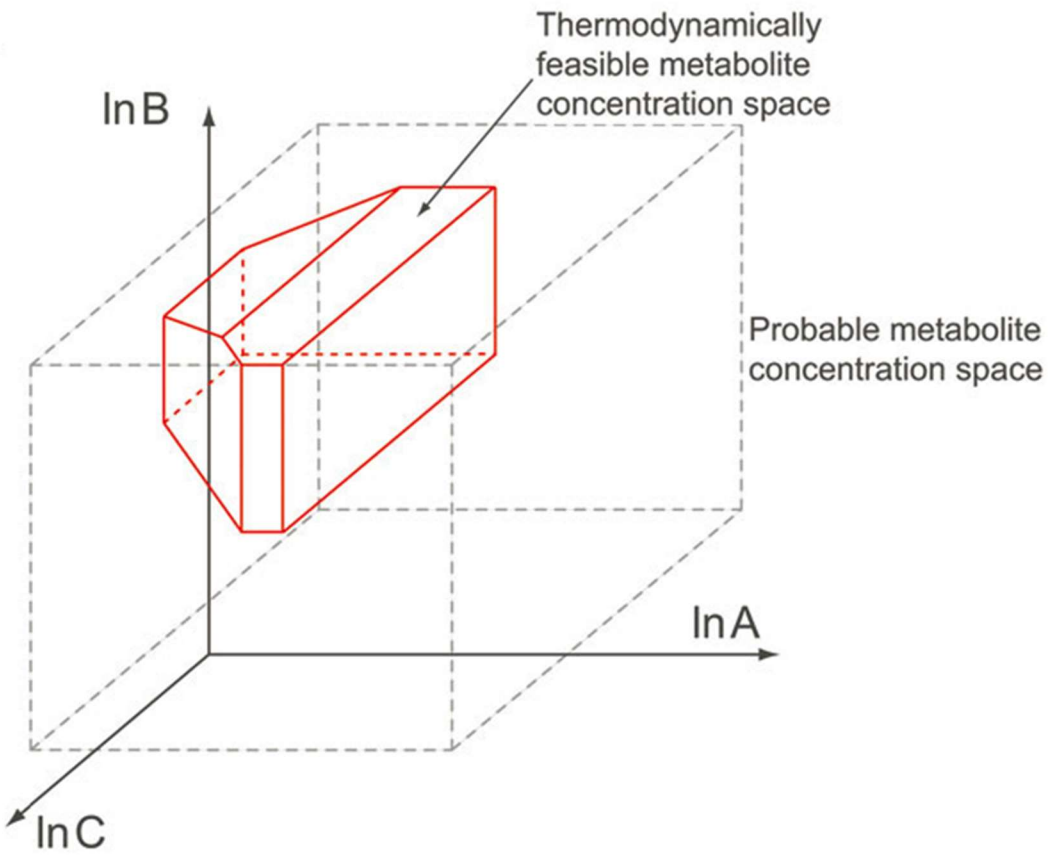

Figure 7:Space of metabolite concentrations within bounds experimentally observed under different physiological conditions (dashed line) and within thermodynamically feasible bounds (solid line)[35].

2.4 Kinetic models and DBTL cycle

2.4.1 Kinetic models

All the constraint-based methods make steady state assumptions and fail to capture the dynamic the metabolic network's dynamic properties. To reveal dynamic properties of reaction networks, kinetics is essential[36]. However, enzyme kinetics traditionally are determined through in vitro experiments which require purification of the enzymes involved. As a result, even for the most studied microorganism S.cerevisiae the vast majority of enzymes lack kinetic parameters. Although experiments are being contacted to determine kinetic properties for enzymes, they will not suffice to describe the full dynamic behavior of a complex metabolic network consisting of hundreds of enzymes.

A different strategy is to try to assume rate laws for each enzymatic reactions and try to predict the kinetic parameters for each reaction. However, reactions such as phosphofructokinase contain 11 different kinetic parameters[37] and the determination of all them proves very difficult. As a result, simplified kinetic laws are assumed for the enzymatic reactions either Michaelis-Menten or mass action kinetics. Most of the time the kinetic models developed predict kinetic parameters with no experimental data available[38-42] or they do not take into consideration known steady state fluxes or concentration data, nor do the ensure thermodynamic constraints[43].

The two main issues that hamper the development of kinetic models are[44]:

a) Uncertainty in metabolite concentrations and thermodynamic displacement. Uncertainties in metabolite measurements and in the estimated thermodynamic properties of reactions (Gibbs free energies of reactions) can impact the conclusions about the displacement of reactions from thermodynamic equilibrium and ultimately the conclusions about the kinetic parameters of the corresponding enzymes.
b) Uncertainty in kinetic properties of enzymes. The lack and uncertainty of information about enzyme kinetics has been acknowledged as the single most important obstacle for developing kinetic models. Uncertainties of this type can be either structural, e.g., incomplete knowledge of kinetic mechanisms, or quantitative, e.g., absent or incomplete knowledge about the values of the kinetic parameters of enzymes.

The Optimization and Assessment of Complex Living Entities (ORACLE) framework tries to bridge the gap between constraint-based methods and kinetic modelling. Using Thermodynamic Flux Analysis to predict potential flux steady states and concentration profiles and then fitting in kinetic parameters to verify the steady state. A population of kinetic models can be generated without sacrificing thermodynamics or steady state fluxes, out of which pruning tests and stability checks will determine potential kinetic models. These models will be used as the basis to construct new metabolic strategies that will lead to the maximization of the desired flux.

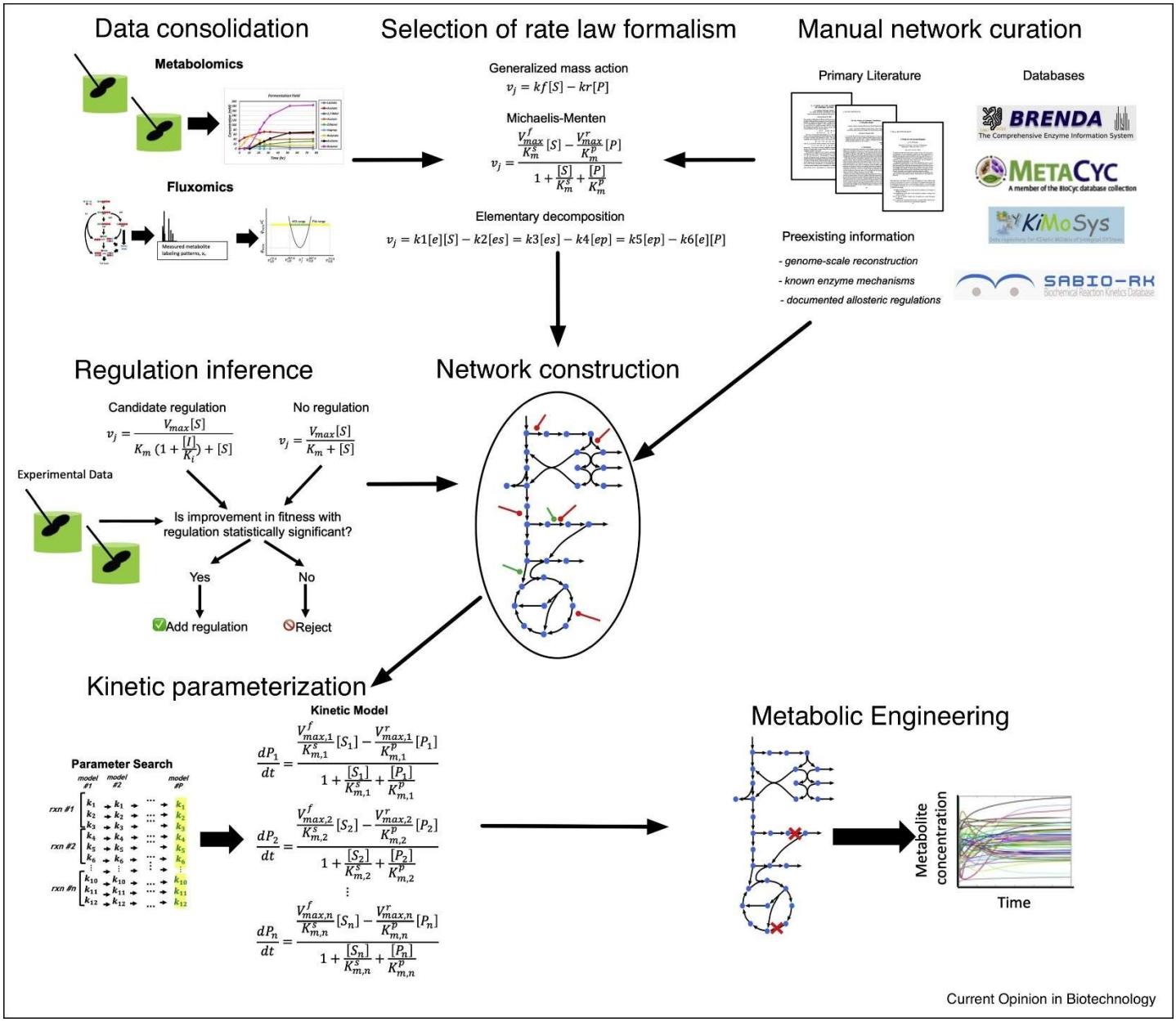

Figure 8: Generalized workflow for metabolic kinetic model construction and use in metabolic engineering[45].

2.4.2 DBTL cycle

The typical design process towards a sustainable bio-production of chemicals follows the iterative Design-Build-Test-Learn (DBTL) cycle. The design involves the selection of the platform organism and the heterologous reactions that need to be expressed to produce the desired chemicals as well as strategies (enzyme upregulations or downregulations) to enhance product yield. The build module involves the genetic transformation and gene editing of the platform organism in accordance to the strategies of the design step. The test module we gather information on the cloning results, omics data and help comprehend cellular behavior. Small batch experiments

showcase the developed strain's desired product yield as well as some key metabolites secretion rates. The learn module takes into account the generated information from the test step and incorporates it into new metabolic strategies to further optimize and increase the desired flux.

The learn module often doesn't take into consideration the kinetic aspect of the cell behavior. Moreover, during the test module the experimental conditions are well monitored and controlled and the yield predictions often fall short during the scaling up of the bioprocess. This uncertainty gap between experimental yield and pilot yield deems bioprocesses untrustworthy and unfavorable for further scaling up. In order to minimize this gap, it is necessary to incorporate large scale kinetic modelling and curating in the learning phase in order to have a clearer picture if we were to move to scaling up process.

Large scale kinetic models not only will offer better metabolic strategies to increase the desired flux during the DBTL cycle but will also work as a basis during the scaling up process. They can be used to simulate the cell behavior in a real batch reactor, with fluctuating conditions while also offering the necessary information for the process optimization. To add to this, secretion rates will be also predicted offering the possibility for downstream separation design and optimization.

Chapter 3. Problem description and methodology

3.1 Problem description and workflow outline

3.1.1 Problem description and main challenges

Typical metabolic engineering strategies to produce efficient cell factories often do not take into account the kinetics and use FBA-related methods which assume that the system is in steady state. The common practice in computer aided strain design involves MILP formulations that reveal metabolic strategies (gene knockouts, reaction additions, gene upregulations or downregulations) that maximize yeast growth coupled with the desired chemical production. These methods may offer a helpful insight and good metabolic strategies but they sometimes fail to take into consideration possible flux bottlenecks due to enzyme saturation or due to the lack of activity from some enzymes. In other words, they fail to predict the dynamic part of a metabolic network and the interplay between stoichiometry, thermodynamics and kinetics.

The development of large-scale metabolic models is a challenging task. The lack of experimental data, the uncertainty of some kinetics parameters available in literature and databases, the uncertainty of the types of mechanisms for every reaction, the errors in metabolomics and fluxomics data are some of the problems[46]. Moreover, for every reaction in the system a rate expression along with values of kinetic parameters are required for a kinetic model. Errors also in the thermodynamic properties hinder the ascertainment of a unique steady-state profile for metabolic fluxes and metabolite concentrations. Taking all this into account it is impossible to find a unique kinetics model which describes the physiology but it is possible to produce a population of models that agree with the physiology and statistical analysis on these models can be used to analyze and predict the metabolic responses in the system[47].

A common practice while developing large-scale kinetic models is to use any type of experimental data such as fluxomics, metabolomics, proteomics, growth rates, uptake rates, secretion rates etc. That way, the problem's uncertainty is reduced and we can get models that are closer to physiology that has been experimentally observed. The alternative approach is to assume ad hoc growth rates and flux directionalities.

The primary challenge will be to produce large-scale kinetic models that are physiologically relevant meaning that their dynamic behavior is close to the experimentally observed one. The secondary challenge is to identify kinetic models that show "robust" behavior. Finally, the population of the kinetics models will be used to target key enzymes and offer metabolic strategies for the increase of flux of the desired product.

3.1.2 Workflow Outline

The proposed workflow outline is based on the Optimization and Risk Analysis of Complex Living Entities (ORACLE) methodology[48-49]. The ORACLE methodology consists of 7 steps:

1) In the first step, the stoichiometry of the system is defined either by biochemical data or genome reconstruction analysis. In this project, a yeast genome scale model is curated and then reduced using redGEM[50] and lumpGEM[51]. Then, the experimental data[52] are integrated into the reduced model as well as the heterologous reactions.
2) In the second step, the solution space from Thermodynamic Flux Balance Analysis is sampled. Metabolite concentrations and reaction fluxes profiles, that agree with thermodynamics, are generated.
3) In the third step, the kinetic parameters for all the reactions are calculated using a Monte Carlo simulation. For every steady state sampled in step 3 the kinetic parameters are calculated to verify that steady state
4) In the fourth step, the produced kinetic models undergo pruning and stability checks. Pruning reduces the models to those that are physiologically relevant and consistent with the experimental data and stability checks identifies the models that show great stability to a wide range of random perturbations.
5) In the fifth step, the stable models are used to calculate flux control coefficients and concentration control coefficients based on a well-established MCA framework[49].
6) In the sixth step, advanced statistical analysis and visualization is performed on the produced populations of control coefficients as well as basins responses of the generated kinetic models.

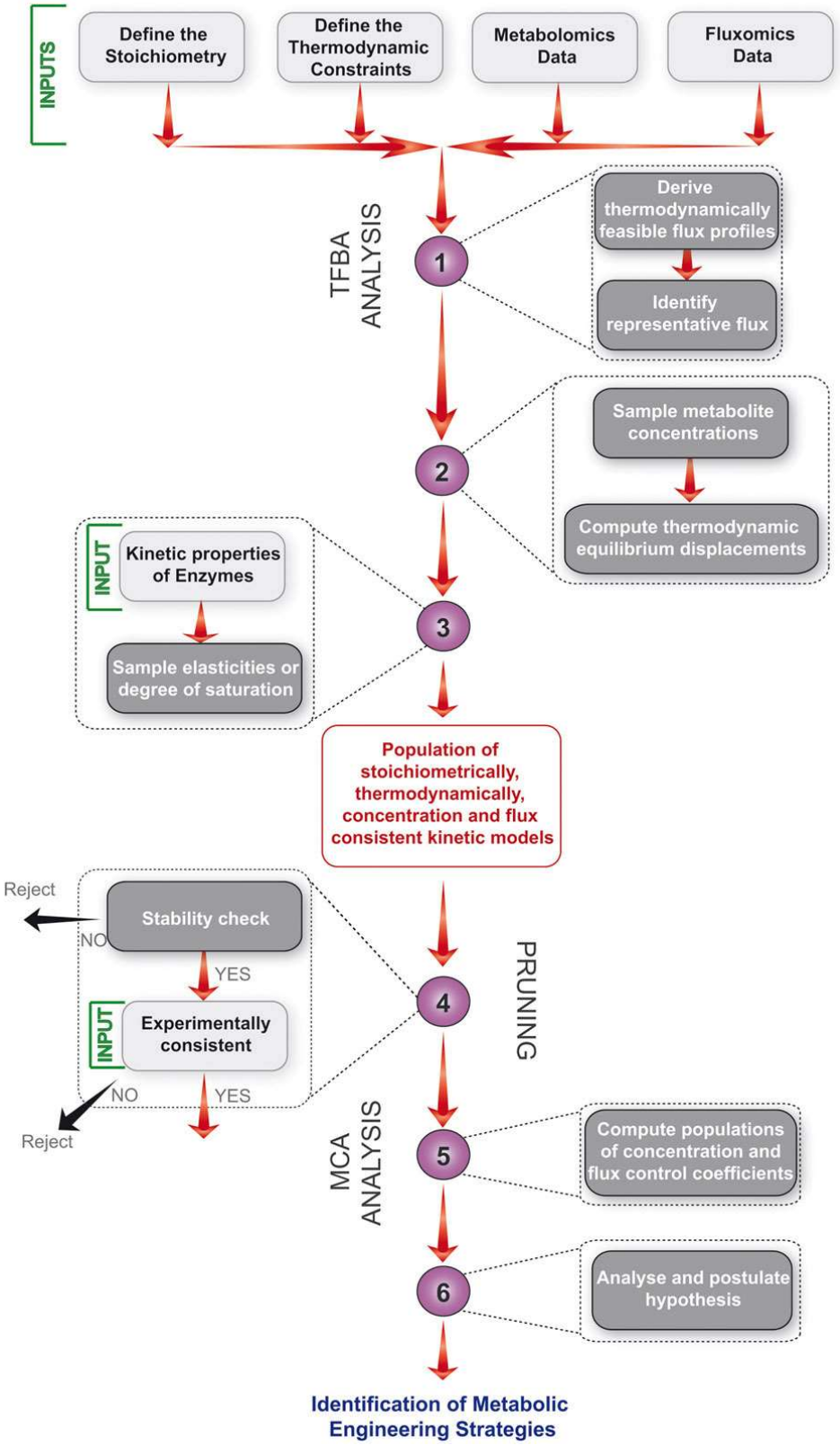

Figure 9:Workflow of the computational procedure for uncertainty analysis of metabolic networks within the ORACLE framework. Light gray boxes denote the integration of available experimental and literature data, whereas the dark gray boxes denote the computation[35]

3.2 Genome Scale Reconstruction, curation and experimental data integration

The first step of the proposed outline is to prepare a functional and practical GEM to conduct the following steps of the analysis. In the vast majority of industrial fermentation applications we utilize one of the so-called platform organisms: *E.coli, S.cerevisiae, C.glutaminicum and A.Niger.* These microorganisms are well-studied and their genome scale metabolic reconstructions are available and thoroughly used. The GEM preparation includes actions concerning the metabolic network and integrating the experimental data.

First of all, the GEM of interest is reduced to specific subsystems of interest. This step is of very important practical value because it substantially reduces the metabolic network, we have to work with thus resulting in a reduction of the necessary computing power. By applying the redGEM[50] and lumpGEM[51] algorithms we can form a core metabolic network on the subsystems of interest without losing paternal GEM characteristics.

On the reduced model we can add the heterologous reactions that will lead to the desired product. The new pathways added are translated to sequential biochemical reactions whose reactants and products are cellular metabolites. The necessary rows for the new metabolites and the necessary columns for the new reactions are added on the stoichiometric matrix along with the corresponding reaction coefficients.

To increase GEM credibility and predictability of the analysis the incorporation of experimental data[52] is a common practice. Fluxomics, metabolomics, physiology, uptake rates, secretion rates, growth can all be used to constrain metabolite concentrations, define reactions direction, add gene deletions, upregulations or downregulations. All this information is valuable in building kinetic models that agree with the observed experimental data and offer a really precise picture of the cellular system and metabolic strategies with solid background.

3.2.1 GEM sampling

In order to develop the large scale metabolic kinetic models, it is necessary to build them around thermodynamically and physiologically feasible steady states. The stoichiometric capacity and thermodynamic constraints imposed on the system through the TFA analysis form a solution space containing all the thermodynamically feasible steady states. In order to maximize the probability of getting close to the steady states that will produce stable and physiologically relevant kinetic models the sampling process is very important.

In order to uniformly sample the solution space an artificial centering hit and run algorithm (ACHR) is used. This algorithm is broadly used to determine flux distributions and metabolite concentrations[53-55]. Flux distribution and metabolite concentration profiles sampled are used in the next step to develop the kinetic models.

3.3 Enzyme kinetics

3.3.1 Modelling and simulation of enzyme kinetics

The rate law can be expressed by the equation below:

$$v = v_{max}\omega\varphi \ (1)$$

where v_{max} is the maximum rate achieved, ω is the mass action expression and ϕ is the saturation component. If we were to take the partial derivative of the rate with respect to the metabolite concentration x we would get:

$$\frac{\partial v}{\partial x} = v_{max}\frac{\partial \omega}{\partial x}\varphi + v_{max}\omega\frac{\partial \varphi}{\partial x} \ (2)$$

Scaling equation (2) by x/v we get:

$$\frac{x}{v}\frac{\partial v}{\partial x} = \frac{\partial ln\omega}{\partial lnx} + \frac{\partial ln\varphi}{\partial lnx} \ (3)$$

The expression on the left-hand side $\frac{x}{v}\frac{\partial v}{\partial x}$ is called enzyme elasticity and it quantifies the interaction between enzyme and metabolite concentration. The expression $\frac{\partial ln\omega}{\partial lnx}$ is defined as the mass action elasticity and the expression $\frac{\partial ln\varphi}{\partial lnx}$ as the saturation elasticity.

The enzyme elasticity is closely related to enzyme saturation meaning that for a high value of elasticity we get low saturation and for a low value we get high saturation[56].

The enzyme elasticity appears to be a sum of mass action elasticity and saturation elasticity:

$$\varepsilon = \varepsilon_m + \varepsilon_s \ (4)$$

3.3.2 Elasticity calculation for a simple uni-bi reaction

In a simple reversible Uni-bi reaction the stoichiometry is as follows:

$$S \leftrightarrow P_1 + P_2 \ (5)$$

Where S is the substrate and P_1 and P_2 the products

The rate expression is:

$$v = \frac{v_{max}\frac{S}{K_{m,S}}\left[1 - \frac{1}{K_{eq}}\frac{P_1P_2}{S}\right]}{1 + \frac{S}{K_{m,S}} + \frac{P_1}{K_{m,P_1}} + \frac{P_2}{K_{m,P_2}}} \ (6)$$

To simplify equation (6) we denote $\Gamma = \frac{1}{K_{eq}}\frac{P_1P_2}{S}$ as the displacement from thermodynamic equilibrium, $\bar{S} = \frac{S}{K_{m,S}}$, $\overline{P_1} = \frac{P_1}{K_{m,P_1}}$, $\overline{P_2} = \frac{P_2}{K_{m,P_2}}$

Using equation (2) and (3) we can get expressions for the enzyme elasticity, the mass action elasticity and the saturation elasticity with respect to every metabolite (S, P_1, P_2):

- $\frac{S}{v}\frac{\partial v}{\partial S} = \frac{1}{1-\Gamma} - \frac{\bar{S}}{1+\bar{S}+\overline{P_1}+\overline{P_2}} \ (7)$

- $\frac{P_1}{v}\frac{\partial v}{\partial P_1} = -\frac{1}{1-\Gamma} - \frac{\overline{P_1}}{1+\bar{S}+\overline{P_1}+\overline{P_2}}$ (8)
- $\frac{P_2}{v}\frac{\partial v}{\partial P_2} = -\frac{1}{1-\Gamma} - \frac{\overline{P_2}}{1+\bar{S}+\overline{P_1}+\overline{P_2}}$ (9)

Table 2:Elasticities for a simple Uni-bi reaction

	Substrate S	**Product P_1**	**Product P_2**
Enzyme elasticity(ε)	$\frac{S}{v}\frac{\partial v}{\partial S}$	$\frac{P_1}{v}\frac{\partial v}{\partial P_1}$	$\frac{P_2}{v}\frac{\partial v}{\partial P_2}$
Mass action elasticity(ε_m)	$\frac{1}{1-\Gamma}$	$-\frac{1}{1-\Gamma}$	$-\frac{1}{1-\Gamma}$
Saturation elasticity(ε_s)	$\frac{\bar{S}}{1+\bar{S}+\overline{P_1}+\overline{P_2}}$	$\frac{\overline{P_1}}{1+\bar{S}+\overline{P_1}+\overline{P_2}}$	$\frac{\overline{P_2}}{1+\bar{S}+\overline{P_1}+\overline{P_2}}$

For the mechanistic enzyme kinetics, we can see that enzyme elasticity is a function of thermodynamic displacement, the stoichiometry of reaction and the metabolites concentrations as well as the saturation constants.

3.3.3 Enzyme saturation

Most enzymatic reactions begin with the binding of regulatory substrate S on the active site of the enzyme E:

$$\mathrm{E} + S \overset{k1}{\rightarrow} ES$$

$$ES \overset{k2}{\rightarrow} E + S \quad (10)$$

The total active site concentration is the sum of the free active site and the site of the substrate/enzyme complex.

$$[E_T] = [E] + [ES] \quad (11)$$

The saturation constant K_m is defined as:

$$K_m = \frac{k_2}{k_1} \quad (12)$$

The degree of saturation can be defined as the ratio of substrate/enzyme complex concentration to total enzyme concentration:

$$\sigma_E = \frac{\frac{S}{K_m}}{\frac{S}{K_m} + 1} \quad (13)$$

and the scaled metabolite concentration can be expressed as a function of degree of saturation:

$$\frac{S}{K_m} = \frac{\sigma_E}{1-\sigma_E} \quad (14)$$

From equation (8) we can deduct that:

- When an enzyme is 0% operating, $\sigma_E = 0$, then $[S] \gg K_m$
- When an enzyme is 100% operating, $\sigma_E = 0$, then $[S] \ll K_m$
- When an enzyme is 50% operating, $\sigma_E = 0$, then $[S] = K_m$

This formulation is very important during the Monte Carlo step because the degree of saturation is well bound between 0 and 1 whereas if we were to sample the saturation constant which is boundless constant it would prove more difficult.

As we consider more complex kinetic mechanisms, it is known that more complex enzyme saturation expressions occur[57] but in the current study equation (9) was used.

3.4 Parameter inference

During the TFA sampling step we generate profiles of steady state fluxes, metabolite concentrations and reaction free energies (v_i, $[X_j]$, $\Delta_r G_j$). Using the information from the steady state solution we will generate kinetic parameters (v_{max}, $K_{m,I}$, K_{eq}) for all the reactions in the system. The generated parameters will have to lead to the corresponding steady state solution and also produce a steady kinetics model.

First of all, for every reaction of the system a mechanistic kinetic expression is added based on biochemical data or in case data is lacking we use ad hoc kinetics.

Table 3:Expressions of mechanistic kinetics used and list of parameters.

Name	Reaction	Rate law expression
Michaelis Menten	$S \rightleftharpoons P$	$v = \frac{V_{max}\frac{[S]}{K_{M,S}}\left(1 - \frac{1}{K_{eq}}\frac{[P]}{[S]}\right)}{1 + \frac{[S]}{K_{M,S}} + \frac{[P]}{K_{M,P}}}$ Parameters: $[V_{max}, K_{M,S}, K_{M,P}, K_{eq}]$
Random Bi-Bi Michaelis Menten	$S_1 + S_2 \rightleftharpoons P_1 + P_2$	$v = \frac{V_{max}\frac{[S_1]}{K_{M,S_1}}\frac{[S_2]}{K_{M,S_2}}\left(1 - \frac{1}{K_{eq}}\frac{[P_1][P_2]}{[S_1][S_2]}\right)}{1 + \frac{[S_1]}{K_{M,S_1}} + \frac{[S_2]}{K_{M,S_2}} + \frac{[P_1]}{K_{M,P_1}} + \frac{[P_2]}{K_{M,P_2}} + \frac{[S_1][S_2]}{K_{M,S_1}K_{I,S_2}} + \frac{[P_1][P_2]}{K_{M,P_2}K_{I,P_1}}}$ Parameters: $[V_{max}, K_{M,S_1}, K_{M,S_2}, K_{I,S_2}K_{M,P_1}, K_{M,P_2}, K_{I,P_1}K_{eq}]$
Generalized Reversible Hill	$S_1 + \cdots + S_N \rightleftharpoons P_1 + \cdots + P_N$	$v = \frac{V_{max}\left(1 - \frac{1}{K_{eq}}\prod_{i=1}^{N}\left(\frac{[P_i]}{[S_i]}\right)\right)\prod_{i=1}^{N}\left(\frac{[S_i]}{K_{M,S_i}}\right)\left(\frac{[S_i]}{K_{M,S_i}} + \frac{[P_i]}{K_{M,P_i}}\right)^{h-1}}{\prod_{i=1}^{N}\left(1 + \left(\frac{[S_i]}{K_{M,S_i}} + \frac{[P_i]}{K_{M,P_i}}\right)^{h}\right)}$ Parameters: $[V_{max}, K_{M,S_1} \ldots K_{M,S_M}, K_{M,P_1} \ldots K_{M,P_M}, K_{eq}, h]$

Uni-Bi Reversible Hill	$S_1 \rightleftharpoons P_1 + P_2$	$v = \frac{V_{max}\left(\frac{[S_1]}{K_{M,S_1}}\right)\left(1 - \frac{1}{K_{eq}}\frac{[P_1][P_2]}{[S_1]}\right)\left(\frac{[S_1]}{K_{M,S_1}} + \frac{[P_1]}{K_{M,P_1}}\frac{[P_2]}{K_{M,P_2}}\right)^{h-1}}{1 + \left(\frac{[S_1]}{K_{M,S_1}} + \frac{[P_1]}{K_{M,P_1}}\right)^{h} + \left(\frac{[S_1]}{K_{M,S_1}} + \frac{[P_2]}{K_{M,P_2}}\right)^{h} + \left(\frac{[S_1]}{K_{M,S_1}} + \frac{[P_1]}{K_{M,P\_1}}\frac{[P_2]}{K_{M,P_2}}\right)^{h} - 2\left(\frac{[S_1]}{K_{M,S_1}}\right)^{h}}$ Parameters: $[V_{max}, K_{M,S_1}, K_{M,P_1}, K_{M,P_2}, K_{eq}, h]$
Bi-Uni Reversible Hill[9]	$S_1 + S_2 \rightleftharpoons P_1$	$v = \frac{V_{max}\left(\frac{[S_1]}{K_{M,S_1}}\frac{[S_2]}{K_{M,S_2}}\right)\left(1 - \frac{1}{K_{eq}}\frac{[P_1]}{[S_1][S_2]}\right)\left(\frac{[S_1]}{K_{M,S_1}}\frac{[S_2]}{K_{M,S_2}} + \frac{[P_1]}{K_{M,P_1}}\right)^{h-1}}{1 + \left(\frac{[S_1]}{K_{M,S_1}} + \frac{[P_1]}{K_{M,P_1}}\right)^{h} + \left(\frac{[S_2]}{K_{M,S_2}} + \frac{[P_1]}{K_{M,P\_1}}\right)^{h} + \left(\frac{[S_1]}{K_{M,S_1}}\frac{[S_2]}{K_{M,S_2}} + \frac{[P_1]}{K_{M,P\_1}}\right)^{h} - 2\left(\frac{[P_1]}{K_{M,P_1}}\right)^{h}}$ Parameters: $[V_{max}, K_{M,S_1}, K_{M,S_2}, K_{M,P_1}, K_{eq}, h]$
Convenience Kinetics	$\alpha_1 S_1 + \cdots + \alpha_M S_M \rightarrow \beta_1 P_1 + \cdots + \beta_N P_N$	$v = \frac{V_{max}\prod_{i=1}^{M}\left(\frac{[S_i]}{K_{M,S_i}}\right)\left(1 - \frac{1}{K_{eq}}\frac{\prod_{j=1}^{N}[P_j]}{\prod_{i=1}^{M}[S_i]}\right)}{\prod_{i=1}^{M}\sum_{m=0}^{\alpha_i}\left(\frac{[S_i]}{K_{M,S_i}}\right)^{m} + \prod_{j=1}^{N}\sum_{m=0}^{\beta_j}\left(\frac{[P_j]}{K_{M,P_j}}\right)^{m} - 1}$ Parameters: $[V_{max}, K_{M,S_1} \ldots K_{M,S_M}, K_{M,P_1} \ldots K_{M,P_N}, K_{eq},]$
Irreversible Michaelis Menten	$S_1 + \cdots + S_M \rightarrow P_1 + \cdots + P_N$	$v = V_{max}\prod_{i=1}^{M}\left(\frac{\frac{[S_i]}{K_{M,S_i}}}{1 + \frac{[S_i]}{K_{M,S_i}}}\right)$ Parameters: $[V_{max}, K_{M,S_0} \ldots K_{M,S_M}]$

The above table contains all the mechanistic kinetic expressions used in the current thesis. The parameters that need to be defined are the v_{max} (maximum rate achieved), the saturation constants K_m which are related to enzyme saturation, the K_{eq} which is related to the ΔG_r and the hill coefficient h which for our problem was equal to 1 for every reaction.

The steps followed to determine these parameters are as follows:

1) The K_{eq} is calculated based on the reaction ΔG_r. The thermodynamic properties needed for this step are calculated during the TFA step and the equilibrium constant is calculated:
$$K_{eq,i} = -RTln(\Delta_r G_i'^{o}) \quad (15),$$
where R is the universal gas constant, T is temperature in and $\Delta_r G_i'^{o}$ is the transformed Gibbs free energy of reaction i calculated during the TFA step.
2) In order to calculate the saturation constants $K_{m,i,j}$ we sample the saturation constants $\sigma_{i,j}$ which are bound $0 \leq \sigma_{i,j} \leq 1$. Then by using equation (13) we calculate the saturation constants.
3) Lastly, the v_{max} parameters are calculated to verify the steady state fluxes.

For each resulting model it is important to test the local stability of the system. In order to do that we calculate the Jacobian matrix:

$$\frac{dX}{dt} = JX + O(X^2),$$

$$j_{i,j} = \frac{\partial}{\partial X_j}\left(\frac{dX_i}{dt}\right) \quad (16)$$

For a metabolic system at steady state, we have:

$$\bar{\bar{N}}\underline{v} = 0 \quad (17)$$

Where $\bar{\bar{N}}$ is the stoichiometric matrix and $\underline{v}$ is the flux vector. The Jacobian matrix can be derived from equation (17)[58-60]:

$$\bar{\bar{J}} = \bar{\bar{N}}\frac{\partial \underline{v}}{\partial \underline{x}} = \bar{\bar{N}}\bar{\bar{V}}\underbrace{\bar{\bar{V}}^{-1}\frac{\partial \underline{v}}{\partial \underline{x}}\bar{\bar{X}}}_{\bar{\bar{E}}}\bar{\bar{X}}^{-1} \overset{(4)}{\Rightarrow} \bar{\bar{J}} = \bar{\bar{N}}\bar{\bar{V}}\bar{\bar{E}}_m\bar{\bar{X}}^{-1} + \bar{\bar{N}}\bar{\bar{V}}\bar{\bar{E}}_s\bar{\bar{X}}^{-1} \quad (18)$$

Where $\bar{\bar{V}}$ is diagonal matrix containing the steady state fluxes, $\bar{\bar{X}}$ is a diagonal matrix containing all the steady state metabolite concentrations, $\bar{\bar{E}}$ is the enzyme elasticity matrix, $\bar{\bar{E}}_m$ is the mass action elasticity matrix and $\bar{\bar{E}}_s$ is the saturation elasticity matrix.

The two elasticity matrixes can be easily calculated from expressions such as those in Table 1. After calculating the Jacobian matrix, we check if the real part of every eigenvalue (λ_i) is negative. If this condition is met then the generated model is stable and the parameter population is stored for further analysis.

3.5 Parameter pruning and stability checks

3.5.1 Parameter pruning

After the parameter inference step, we have generated a large number of kinetic models containing fluxes, concentrations and parameters. However, the majority of those models appear not to be physiologically relevant and would result in uncertain predictions. As the first step of the parameter pruning, we check if the system response to a small metabolite concentration perturbation is within physiological bounds.

For a small perturbation the response of metabolite i concentration will be given by the equation:

$$[X_i] - [X_i]_{steady} = \left([X_i]_{t=0} - [X_i]_{steady}\right)\exp\left(-\frac{t}{\tau_i} + i\omega_i t\right) \quad (19)$$

Where $\tau_i = abs\left(\frac{1}{Re(\lambda_i)}\right)$ and $\omega_i = Im(\lambda_i)$. The first term $-\frac{t}{\tau_i}$ inside the exponential represents the exponential decay and second term $i\omega_i t$ represents the harmonic oscillations. By ignoring the term for the harmonic oscillations, the dynamic response of the system would be an exponential decay reaching the reference steady state and the time constant of the system would be τ_i.

We define two bounds for the time constants:

- The upper bound is $\tau_{max} = \frac{t_d}{4}$,where t_d is the doubling time of the cell. The upper bound ensures that the system has reached quasi steady state meaning that the changes in metabolite concentrations levels are so small that could be considered constant.

- The lower bound is $\tau_{min} = k_{diff}^{max}[E_{min}] = 10^{-7}$. This ensures that metabolism is slower than physics.

The physiological relevant models are those that all the time constant values are within those bounds.

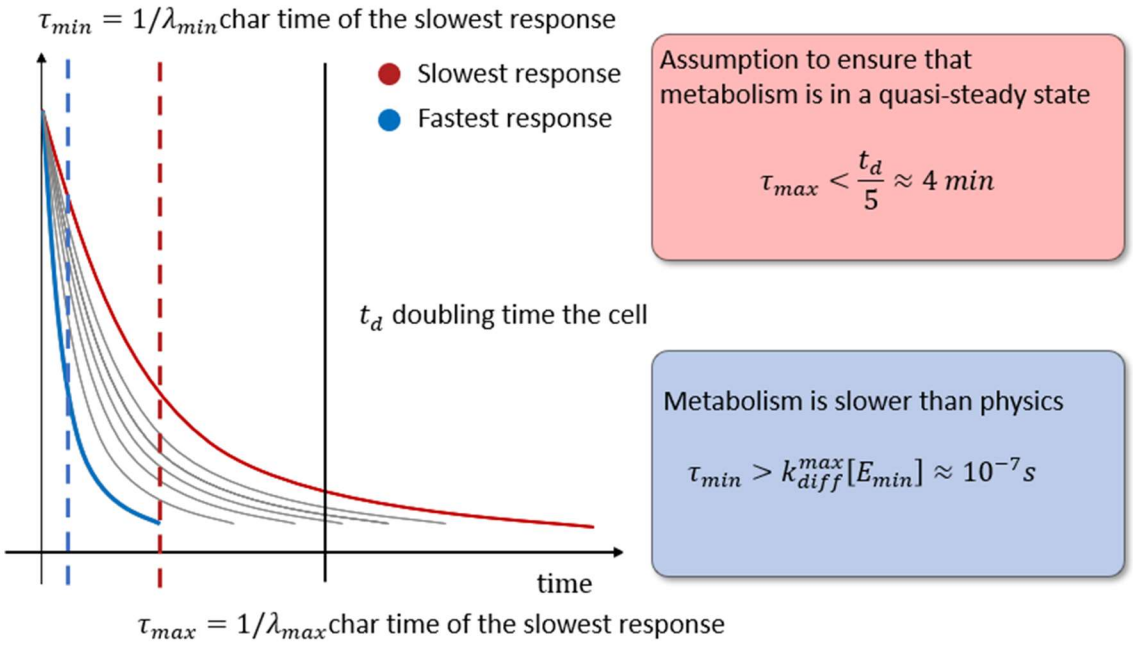

Figure 10:The model's response must be within physiological bounds.

3.5.2 Assessing robust model behavior (Basins)

The models surviving the pruning step are further analyzed to identify models that show great stability around their steady state. We also tested the models for different enzyme saturation states. To make this analysis we performed two sets of random perturbations on the metabolite concentrations values. The first set of perturbations was on a short range and the second set was on a wider range.

Table 4: Ranges of perturbations.

Short range	Wide range
$0.8[X_{ref}] \leq [X] \leq 1.2[X_{ref}]$	$0.5[X_{ref}] \leq [X] \leq 2[X_{ref}]$

For every metabolite in the system, we imposed random initial conditions on the concentration values within the defined ranges, thus creating a system of ordinary differential equations, an initial value problem. Solving the system provided us with the trajectories of responses for the metabolites of the models. We further classified the trajectories based on the distance from the reference steady state. The distance was calculated as the norm:

$$Distance = \left| \frac{[X_i] - [X_i]_{ref}}{[X_i]_{ref}} \right| \quad (20)$$

After a lot of time ($t \to \infty$), the trajectories indicated three different scenarios:

- Metabolism reached the same steady state. Every metabolite concentration returned to the reference steady state.
- Metabolism reached another steady state. A number of metabolites ended up in a different steady state concentration value.
- Metabolism reached a pathological state. Some metabolite concentrations did not converge to a steady state but rather escaped further increasing with the pass of time.

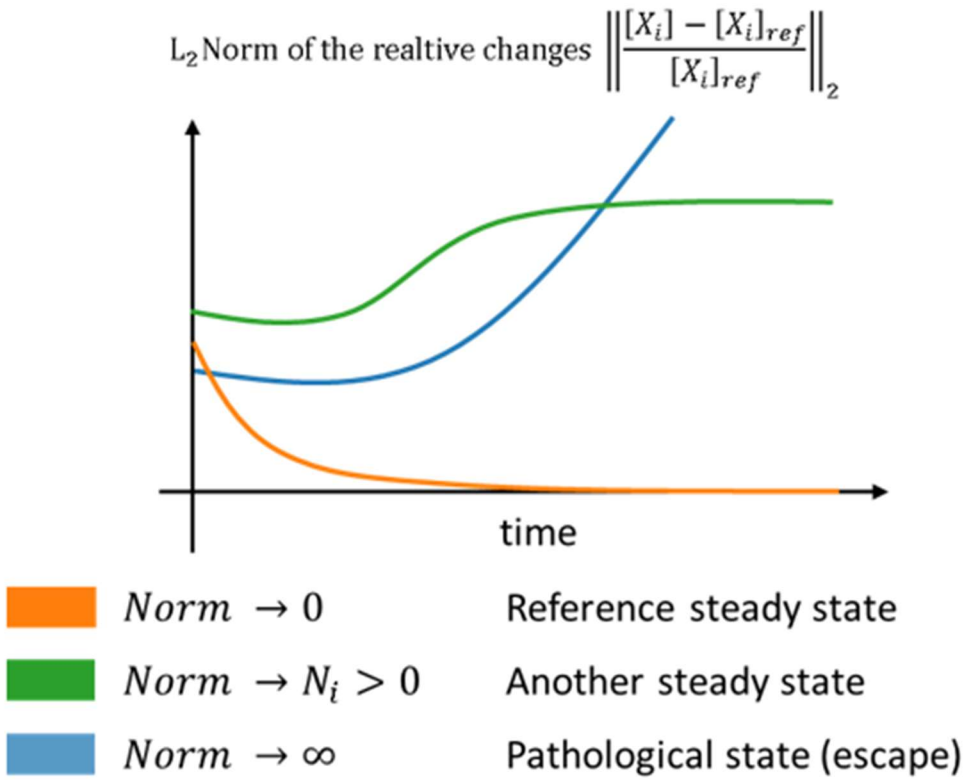

Figure 11:The three different scenarios for the metabolite concentration trajectories

3.5.3 Clustering trajectories

To further analyze the generated kinetic models, we performed clustering on the trajectories produced during the basins step. First of all, we classified the basins responses only by the three aforementioned states without clustering. This way, we screened the models that reached new steady states and used those for the clustering.

For the clustering process we used the models whose trajectories reached another steady state. As an input we used the metabolite concentration vector after a long time ($t \to \infty$) and also added the reference steady states provided during the TFA step. This way we were able to identify if

different models reached same steady states in case, they didn't reach the reference steady. For the clustering we used the kmeans++[61] algorithm embedded in sklearn in python. KMeans++ is an iterative centroid based clustering technique in which n data points are arranged to k clusters. Each observation belongs to the cluster with the nearest mean. In order to define the number of clusters we used the elbow method where the distance of clusters is plotted against the number of clusters.

3.6 Metabolic Control Analysis (MCA)

Having performed the pruning and stability check we identify the kinetic models that are physiologically relevant and show great stability. These models will be used to calculate the flux control coefficients for the desired reaction or reactions.

The flux control coefficients C_p^v and the concentration control coefficients C_p^x are defined as the fractional change of metabolic fluxes and metabolite concentration, respectively, in response to fractional changes in parameter values. Having linearized and scaled the system around the steady state[62-63] we can derive the expressions for the control coefficients:

$$C_p^x = -(N_R V E_i + N_R V E_d Q_i)^{-1}[N_R V \Pi_m \vdots N_R V \Pi_e \vdots N_R V \Pi_s],$$

$$C_p^v = (E_i + E_d Q_i) C_p^{xi} + [[\Pi_m \vdots \Pi_e \vdots \Pi_s] \quad (21)$$

Where V is a diagonal matrix with the steady state fluxes, N_R is the reduced stoichiometric matrix, E_i and E_d are the elasticity matrices with respect to independent and dependent metabolites respectively, Π_m, Π_e and Π_s are the matrices of the elasticities with respect to parameters, Q_i is a weight matrix that represents the relative abundance of dependent metabolites with respect to independent. A weight matrix Q_m is defined as the abundance of dependent metabolites with the respect to the levels of their corresponding moieties, which leads to the expression for the matrix of moieties elasticities with respect to parameters, Π_m:

$$\Pi_m = E_d Q_m \quad (22)$$

Symbol	Name	Dimension
N	Stoichiometry	$m \times (n+n')$
N_R	Reduced stoichiometry	$m_0 \times (n+n')$
p	Overall parameters	$(n+m'+s)\times 1$
p_e	Enzyme activity parameters	$n \times 1$
p_m	Conserved moiety concentrations	$m'\times 1$
p_s	Other system parameters	$s \times 1$
v	Metabolic fluxes	$(n+n')\times 1$
V	Steady state fluxes	$(n+n')\times(n+n')$
x	Metabolite concentrations	$m \times 1$
x_d	Dependent metabolite concentrations	$m'\times 1$
x_i	Independent metabolite concentrations	$m_0 \times 1$
X_i	Independent metabolite concentrations	$m_0 \times m_0$

Symbol	Name	Definition	Dimension
$C_p^{x_i}$	Concentration control coefficients	$C_p^{x_i} = \frac{d\ln x_i}{d\ln p}$	$m_0 \times (n+m'+s)$
C_p^v	Flux control coefficients	$C_p^v = \frac{d\ln v}{d\ln p}$	$(n+n')\times(n+m'+s)$
E_i	Independent concentration elasticities	$E_i = \frac{\partial\ln v}{\partial\ln x_i}$	$(n+n')\times m_0$
E_d	Dependent concentration elasticities	$E_d = \frac{\partial\ln v}{\partial\ln x_d}$	$(n+n')\times m'$
Π_e	Enzyme parameter elasticities	$\Pi_e = \frac{\partial\ln v}{\partial\ln p_e}$	$(n+n')\times n$
Π_m	Moiety parameter elasticities	$\Pi_m = \frac{\partial\ln v}{\partial\ln p_m}$	$(n+n')\times m'$
Π_s	Other parameter elasticities	$\Pi_s = \frac{\partial\ln v}{\partial\ln p_s}$	$(n+n')\times s$
Q_i	Relative weights	$Q_i = \frac{\partial\ln x_d}{\partial\ln x_i}$	$m'\times m_0$
Q_m	Absolute weights	$Q_m = \frac{\partial\ln x_d}{\partial\ln p_m}$	$m'\times m'$

Note:
(1) m : number of metabolites
(2) n : number of net fluxes
(3) n' : number of reversible fluxes
(4) m_0 : number of independent metabolites ($m_0 = rank(N)$)
(5) m' : number of dependent metabolites, number of conserved moieties ($m' = m - m_0$)
(6) s : number of other parameters

Figure 12:MCA nomenclature[49]

The flux control coefficients will provide a good insight on which enzymes affect the desired reaction flux. The enzymes identified by MCA will be put to the test using the generated models. Enzyme perturbations will be performed to quantify and evaluate the flux change predicted by the MCA framework.

Chapter 4. Case study: Muconic acid producing yeast

4.1 GEM model preparation

4.1.1 Model reduction

The genome scale model for Saccharomyces Cerevisiae used in the current thesis was the yeast8 model[64]. The reduction of the model was not done on the current thesis and the reduced model was provided. The reduced model contains 5 subsystems that form our metabolic network:

- Glycolysis
- Pentose Phosphate Pathway
- TCA cycle
- Oxidative Phosphorylation

- Shikimate pathway (aromatic aminoacids)

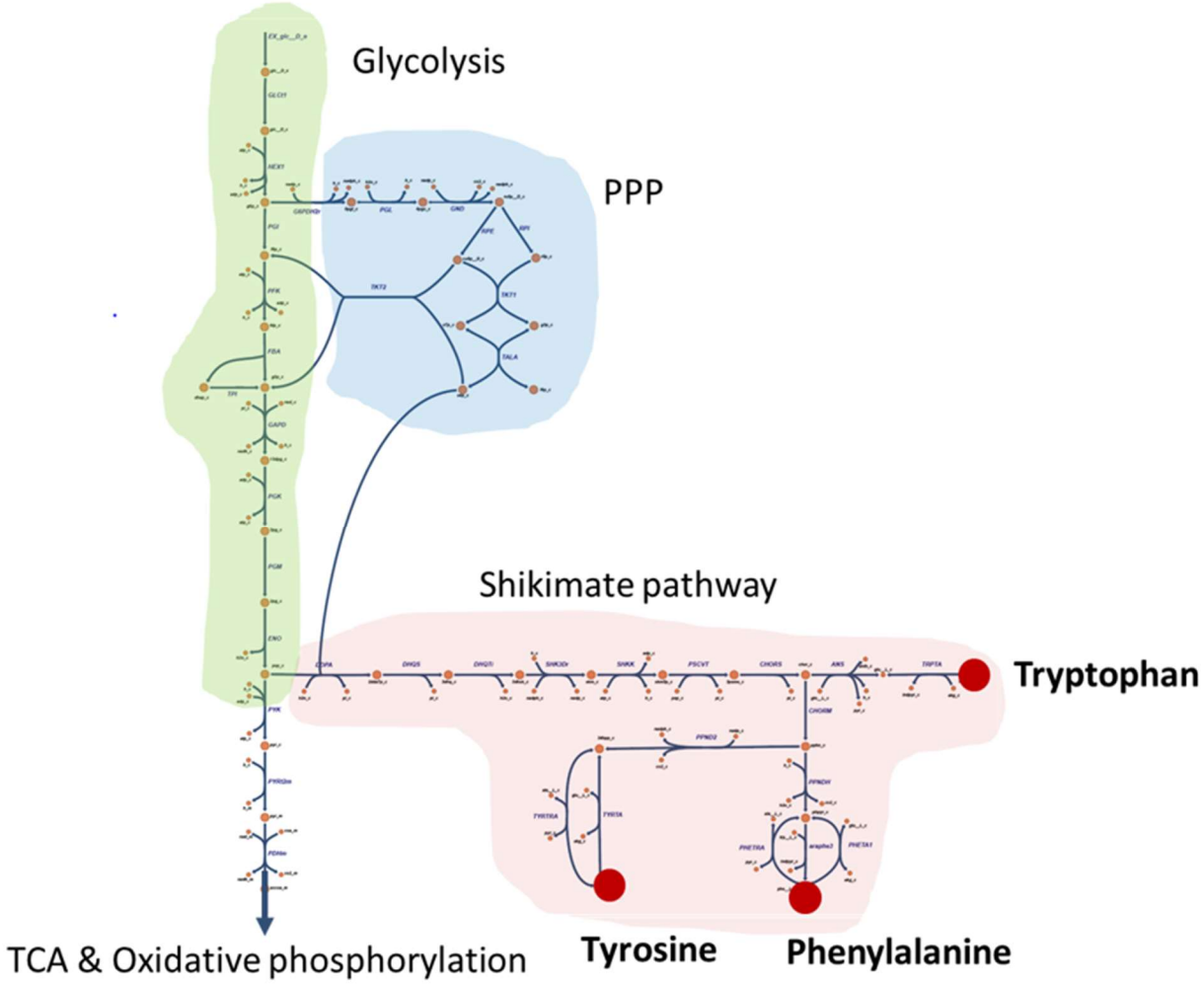

Figure 13:Reduced yeast metabolic pathway

The reduced yeast model has 226 metabolites, that makes a total of 226 mass balances. Out of those, 49 are extracellular metabolites. A total of 308 reactions describes the metabolic network out of which 121 are transport reactions, 183 are enzymatic and 1 lumped biomass that describes biomass production and subsequently yeast growth. The yeast cell is divided into 9 different compartments, each with a different volume and different reactions happening there:

Table 5: Yeast cell compartments and their volume

Compartment	Volume(µm^3)
Cell	42
Cytosol (_c)	29.4
Mitochondria (_m)	0.0378
Inner Mitochondria (_i)	0.0042

Golgi (_g)	4.2
Golgi membrane (_gm)	42
Endoplasmic reticulum (_er)	0.84
Endoplasmic reticulum membrane (_erm)	42
Vacuole (_v)	2.94
Cell envelope (_ce)	42

4.1.2 Experimental data integration

First of all, we added the heterologous reactions for the muconic acid production pathway. According to the experimental data[52] the pathway used was the shunting of the shikimate pathway. We assumed that the reactions added took place in the cytosol.

Table 6: Hetetologous reactions added to the reduced model

Enzyme	Reaction
PaAroz	3dhs_c ⇔ pca_c + h2o_m
KpAroY	pca_c ⇔ catechol_c +co2_c
CaCatA	catechol_c + o2_c ⇔ ccm_c
ccmt2p	ccm_c⇔ ccm_e
pca2tp	pca_c⇔pca_e

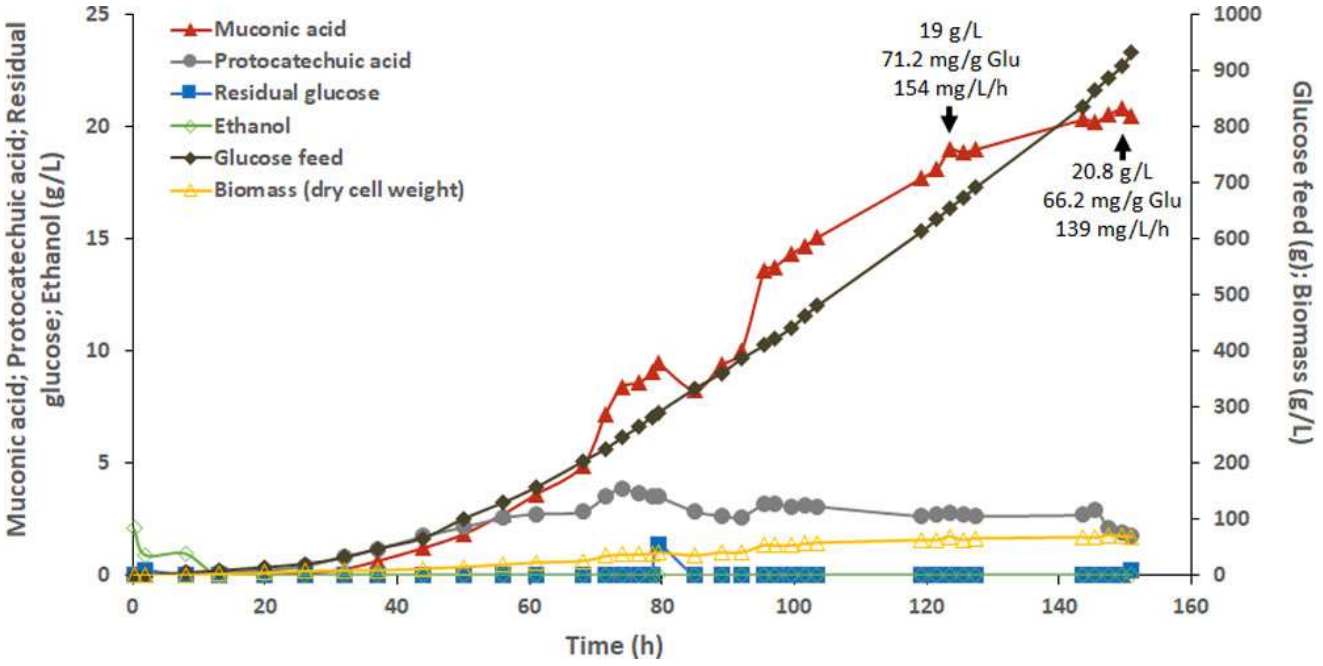

Figure 14:Fermentation data used to constrain the model

The metabolomics from the fitting of the fermentation data shown in Figure 2 were used to constrain the model. Also, gene KOS were expressed in the form of setting the lower and upper bound of the corresponding reaction to zero. That way it is assured that the reaction flux during

the TFA step will be zero. Moreover, reaction directionalities will be defined either from literature data. The extracellular space in this formulation is considered to be equivalent to reactor space, meaning that the extracellular concentrations is the reactor concentration.

1) Glucose extracellular concentration was constrained between 9.8M≤ glc_c ≤10.2M
2) Extracellular fluxes for glucose, muconic acid and prtocatechuic acid are integrated:

Table 7: Extracellular fluxes bounds

Reaction	Lower bound	Upper bound
EX_glc__D_e	-0.55	0
EX_ccm_e	0.028	0.0308
EX_pca_e	0.008	0.0088

3) Define reaction directionalities:
   - Glycolysis is modeled to operate only forward. The reactions FBA,PGI,PGM are set to forward
   - Assume diffusion of passive transporters (akg_m, mal_m) as they were not observed in the medium. Transport reactions AKGt, MALt set to backward.

Table 8: Other fermentation data used

Ratio gdw to gww	0.32
Density	1200 g/l
Ph(extracellular)	6
Minimum growth	0.03h^{-1}
Division time	ln2/minimum_growth=23.1h

4.2 Thermodynamic Flux analysis Sampling

Having constrained the reduce model based on the fermentation data provided we move on to the sampling step. During the sampling process we had to first get the metabolite concentration close to the fluxes calculated. This is an important step, because during the pruning step we need the time constants for every metabolite to be within some physiological bounds. In order to efficiently get the metabolite concentration close to the steady state fluxes, we followed the following steps:

1) First, we took 500 samples from the TFA solution space. We could continue with the parameter inference step, calculate the time constants and proceed with the metabolite concentration curation but that would need a lot of computational power and time.
2) Then we calculate the turnover for the metabolites in the system. Turnover is the following expression:

$$turnover = \frac{\sum v_{prod,i}}{[X]_i}$$

Where $v_{prod,I}$ are all the fluxes that produce metabolite i and $[X]_i$ is the metabolite concentration. The turnover index is somewhat analogous to the time constants of

the metabolites mass balances. The median, maximum and minimum of the 500 samples were calculated and used to indicate which metabolites had small turnovers.

3) The concentrations of the metabolites with low turnovers were further constrained in order to increase the turnover index and subsequently their time constant.

The three steps above were repeated until the turnover indexes for all metabolites showed values greater than 0.1.

4.3 Kinetics Models generation

4.3.1 Pruning step

In order for a kinetic model to be physiologically relevant we demand that the maximal eigenvalue:

$$\lambda_{max} < -3\frac{t_d}{ln2} => \lambda_{max} < -0.12$$

At first, having not done the turnover procedure we sampled 100 flux and concentration profiles from TFA. We then fitted 100 kinetic parameter sets for every TFA sample and ended up with 1000 kinetic models. However, none was physiologically relevant as the histogram below hints.

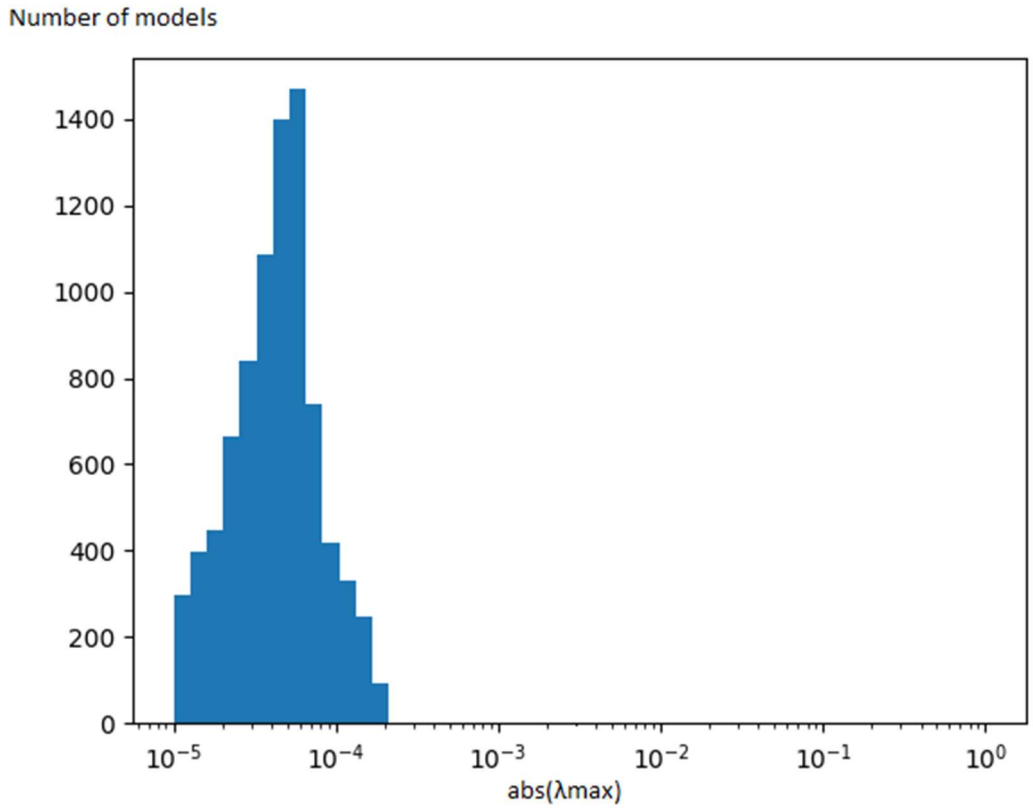

Figure 15:Histogram of maximal eigenvalues of all 100000 models before the turnover step

As it is visible from Figure 3, we were 3 points of magnitude below the cutoff point. After the turnover procedure we generated 100000 kinetic models (1000TFA*100Kinetic_sets) and ended

up with only 56 of those passing the pruning step. As the number of pruned models was so low, we screened the tfa samples to those that generated models with maximal eigenvalues close to the cutoff value. In the end, 47 tfa samples were used and 500 kinetic parameter sets were generated for each sample, thus resulting in 23500 models. Out of those, 366 models were only physiologically relevant, a percentage of 1.56%.

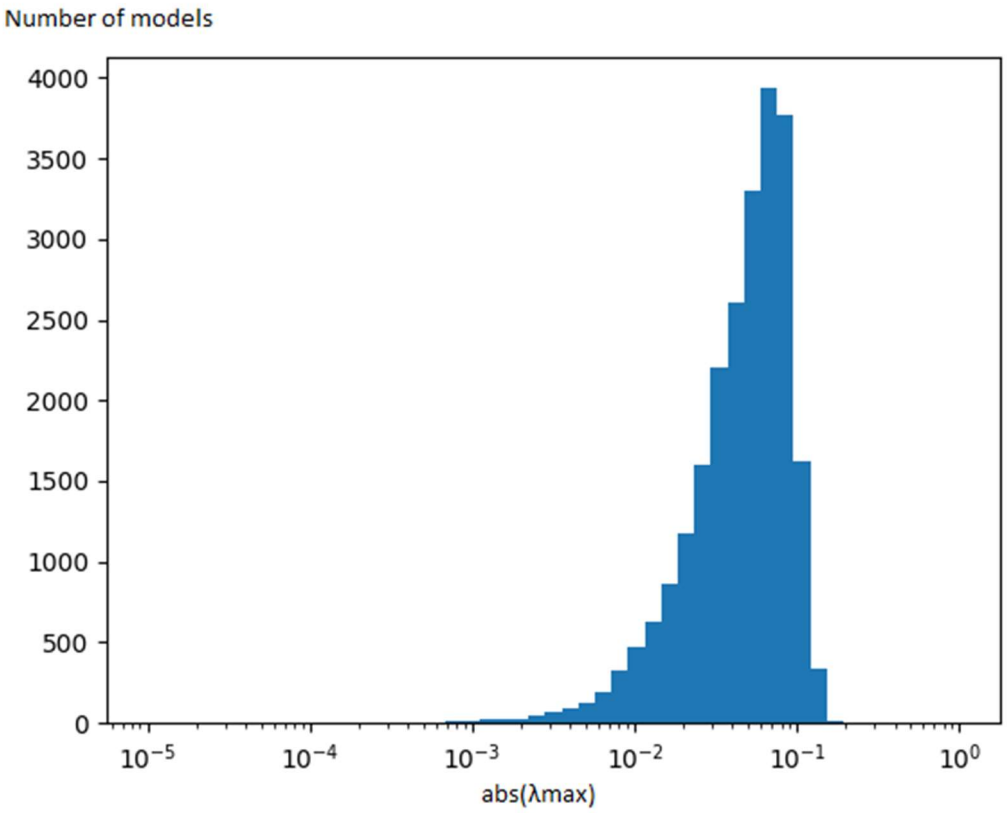

Figure 16:Histogram of maximal eigenvalues of all 23500 models. It is visible that we are closer to the cutoff value

4.3.2 Basins Distributions

To assess the stability of the generated and pruned models we performed the sets of random perturbations. Then we classified the trajectories of the models based on the three different outcomes: reference steady state, another steady state or escape. We will show the basins distributions for both ranges of perturbations. We will start with the models that showed great stability in the short range and juxtapose the basins distribution for the same models on the wide range.

The systems of ordinary equations were solved for a time span of 0min-100min. Then the timeframe used to evaluate the distributions was for t=23min which is the doubling time of the cell.

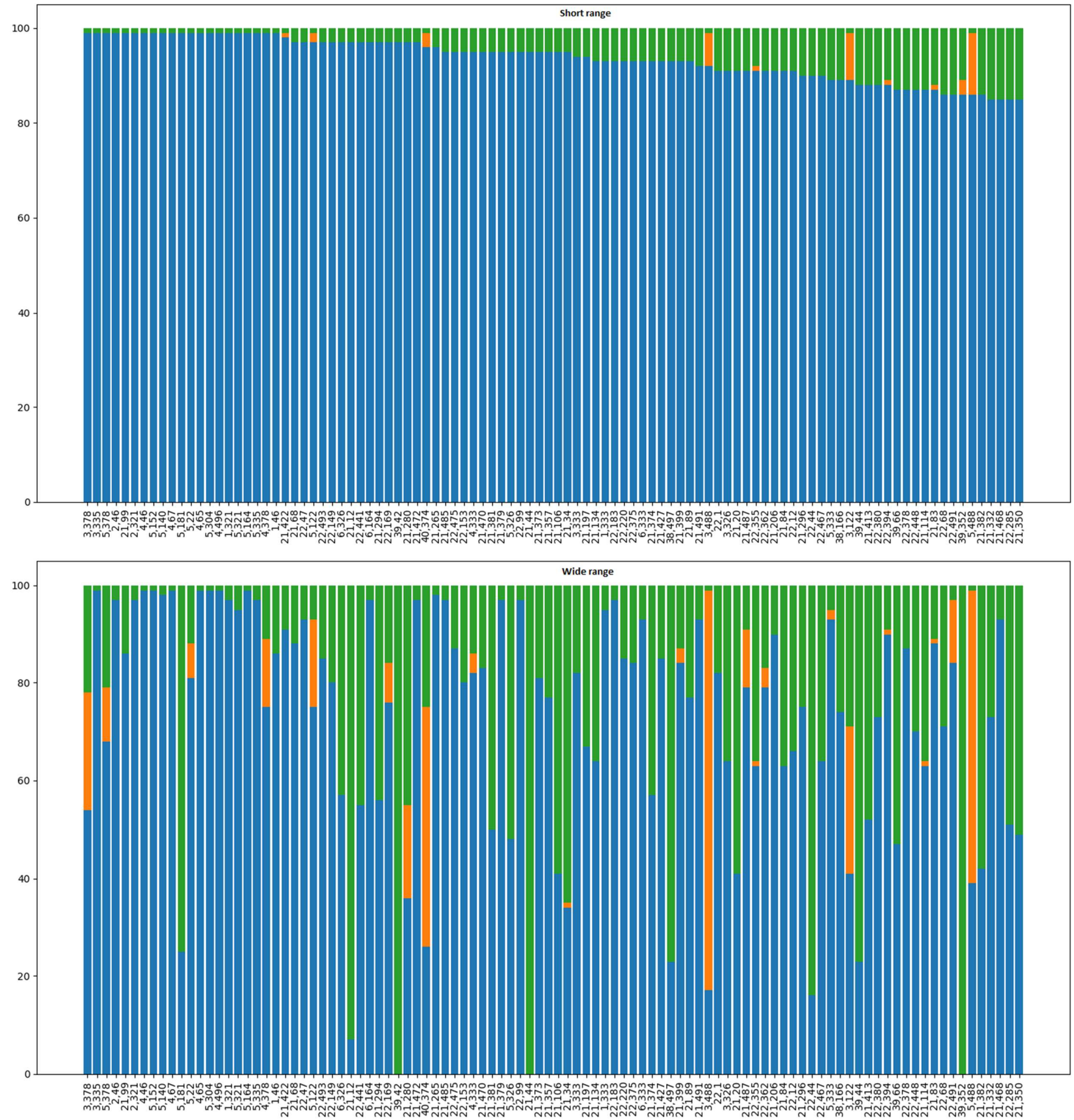

Figure 17: Basins distributions of models 0-100. Blue bar is for the reference steady, green bar is for escape and orange bar is other steady state

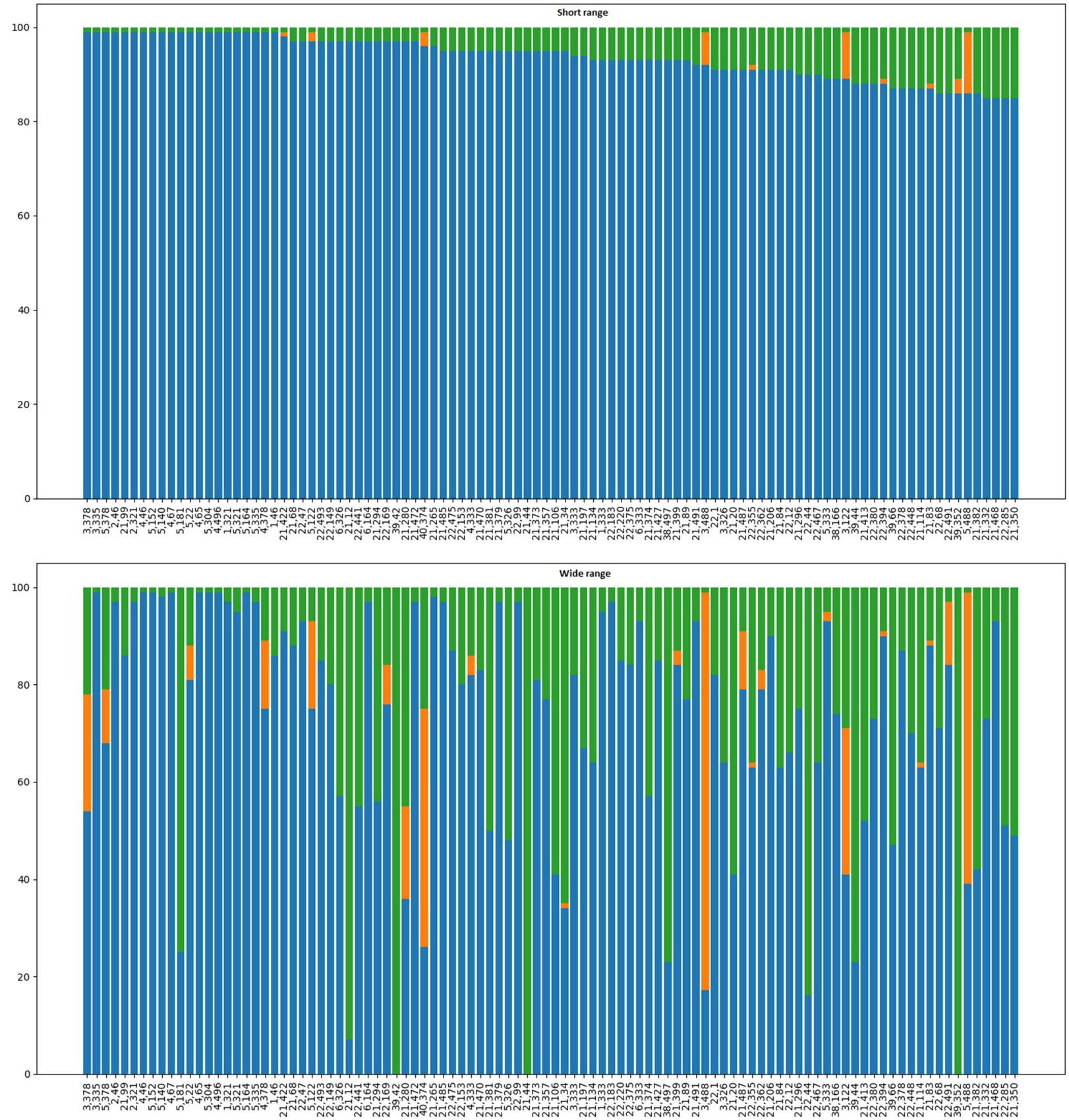

Figure 18: Basins distributions of models 100-200

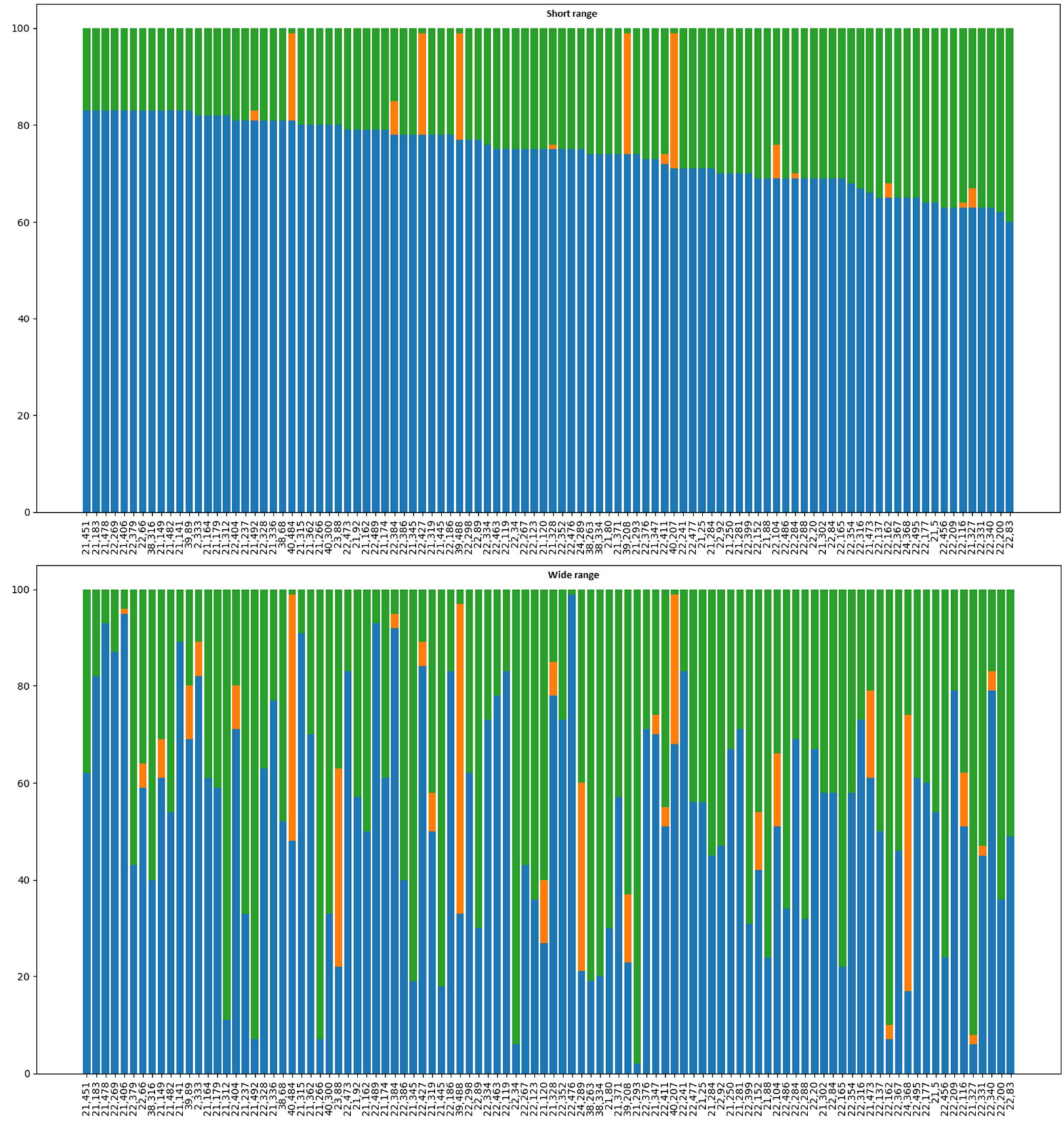

Figure 19: Basins distributions of models 200-300

Figure 20: Basins distributions of models 300-366

Some good observations that can be made from the basins distributions for both ranges are:

- Models that show great stability in the short range are more likely to show great stability in the wide range but that is not guaranteed. For example, model 39,107 had 99/100 perturbations return to reference steady state for the short range but scored 0/100 for the wide range.
- On the wide range, models reached new steady states more frequently than on the short range. This is normal as on a wider specter of initial concentrations metabolites can reach new steady state concentrations.
- On the wide range, models reached a pathological state (escaped) far more often than on the short range. The further away we go from the reference steady state the more probable that some metabolite concentrations don't reach a steady state but rather continue to increase to infinity.
- Models that reached new steady states on the short range of perturbations also reached new steady states on the wide range. Further clustering analysis will show if the new steadies coincide.
- Unstable models in the short range also showed great instability on the wide range with some exceptions (models 40,66 and 40,89).

4.3.3 Basins trajectories clustering

To delve deeper into the understanding of these systems we clustered the steady states observed throughout the basins step. The cluster points will provide information on metabolites that reached new steady states. For the clustering we used the basins results for the models on the wide range that reached a new steady. To define the number of clusters we used the elbow method.

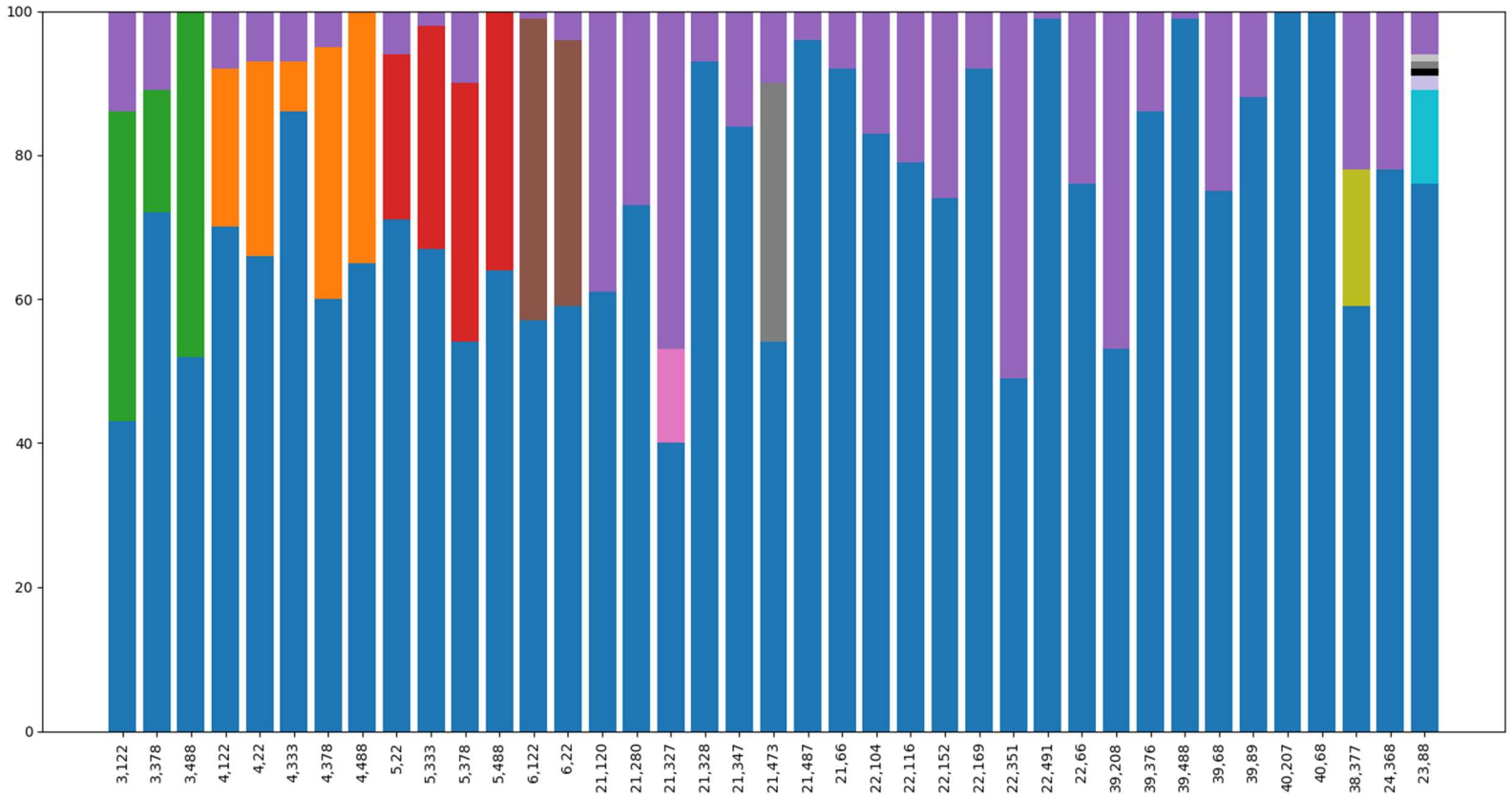

Figure 21:Distribution of basins for models that reached new steady states on the wide range. Blue bars represent the reference steady state, purple bars represent escape and all the other colors represent new steady states.

From Figure 21 we can clearly see that models built around the same steady state reached the same alternative steady state (the case for the models built around steady states with index 3,4,5,6). Other models based on the same steady state (index 21) reached different alternative steady states. And also, the case for the model 23,88 which reached 5 different steady states. However, three of them should be included as in the escape cluster as they reached unnaturally big concentrations.

In the matrix below we will show the distance of the 5 metabolites that are the farthest from the reference steady for every cluster.

Table 9: Other steady state characteristics

Cluster no	Met 1	Norm	Met 2	Norm	Met 3	Norm	Met 4	Norm	Met 5	Norm
3(green)	pram_c	0.14	ile__L_c	0.070	ile__L_m	0.058	pi_erm	0.039	imp_c	0.036
4(orange)	pram_c	0.061	ile__L_c	0.035	imp_c	0.031	ile__L_m	0.027	s_0834_m	0.022
5(red)	pram_c	0.066	imp_c	0.033	pi_erm	0.025	cbc_c	0.021	s_0834_m	0.020
6(brown)	ile__L_c	0.14	ile__L_m	0.11	s_0834_m	0.075	3mob_c	0.04	imp_c	0.036
21,1(pink)	3mob_c	0.12	val__L_c	0.11	3mob_m	0.044	tyr__L_c	0.038	lald__L_c	0.024
21,2(grey)	val__L_c	936	ile__L_c	750	tyr__L_c	665	ile__L_m	292	pram_c	170
38	pi_erm	0.051	cbc_c	0.037	cbasp_c	0.030	gmp_c	0.020	gtp_c	0.020
23,1(light blue)	val__L_c	2.2e7	3mob_c	4.3e5	3mob_m	487	3c3hmp_m	4.9	3c3hmp_c	4.6
23,2	akg_m	5.1e12	s_0834_m	2.2e10	val__L_c	2.3e8	indpyr_C	4.8e5	trp__L_c	1.4e5
23,3	akg_m	5e12	nadp_er	3.6e10	s_0834_m	2.1e10	eig3p_c	1.4e10	val__L_c	2.3e8
23,4	akg_m	5.1e12	s_0834_m	2.2e10	eig3p_c	1.4e10	val__L_c	2.3e8	prlp_c	1.7e6

4.3.4 Models with great stability

Having done the basins analysis we can identify models that show robust behavior and great stability. We define as stability the fraction of the number of perturbations that returned to the reference steady state to the total number of perturbations:

$$S = \frac{n_{reference\_steady}}{n_{sets}}$$

The other classes of trajectories are defined as unstable as it is not physiologically logical for the metabolites concentrations to either reach other steady state values in case of a small perturbation or escape to very large values.

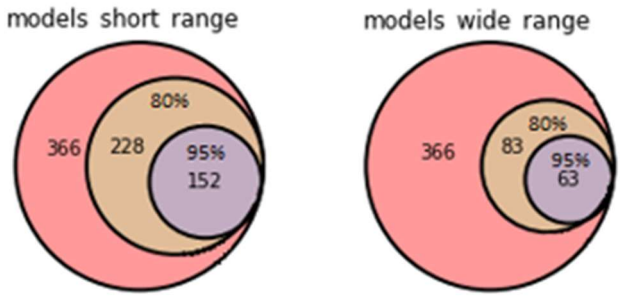

Figure 22:Venn diagrams of models for two different kind of stabilities 95% and 80%.

It is logical that for the shorter range of perturbations more models scored higher stability than for the wide range.

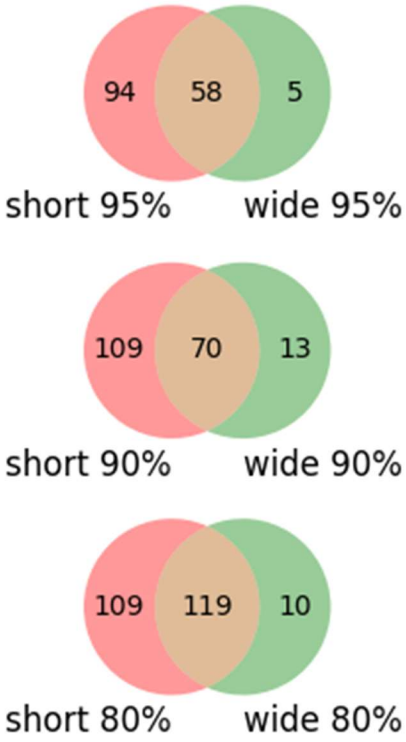

Figure 23:Venn diagrams for both ranges of perturbations and three different stability scores. Intersection of models also illustrated.

Models with high stability in the wide range are very likely to have high stability in the shorter range as well. As stable models are selected those that had a stability score of 90% for both ranges of perturbations. A total of 70 kinetic models will be used to calculate the flux control coefficients for muconic acid production.

4.3.5 TFA samples and stability

To further illustrate the importance of the steady state flux and concentration profile prediction we present the percentile of fast (physiologically relevant models) and stable models for every tfa sample used to generate the kinetic models. For every tfa sample we generated and fitted 500 random kinetic parameter sets.

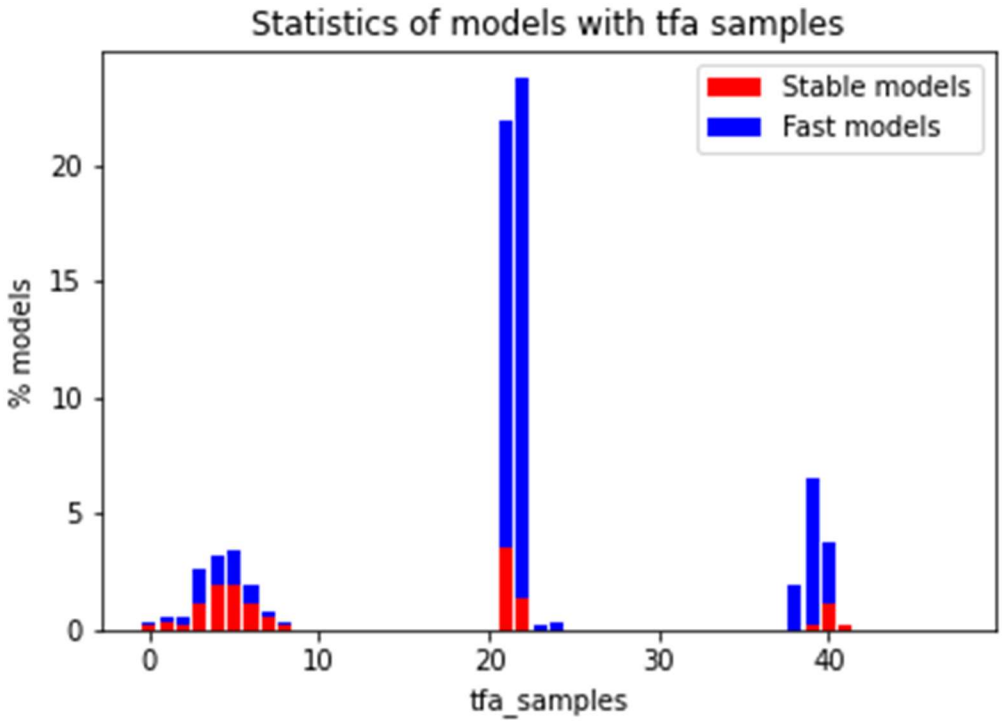

Figure 24:Physiological and stable models generated for every tfa sample

It is visible that while some samples produced many fast models, they didn't necessarily produce stable models.

4.3.6 Decision tree analysis on key parameters

After the stability analysis we have deducted that out of the 366 kinetic models 70 had high stability scores. In order to evaluate and find the kinetic parameters that deem a kinetic model unstable or stable we performed classification using a Decision Tree Classification Algorithm. As the input data for the kinetic parameters, we used:

1) The enzyme saturation values. The saturation values are well bound between 0 and 1 and they didn't need further normalization
2) The thermodynamic displacement(Γ) values. The Γ values are well bound as well between 0 and 1.

The v_{max} values were not used as they would require further normalization and also, they are dependent variables. As a training set for the algorithm, we used all 366 stable kinetic models as we wanted to extract information on the different kinetic parameter values that affect stability rather than make predictions. We generated two decision trees, one for stability score 90% and one for stability score 80%.

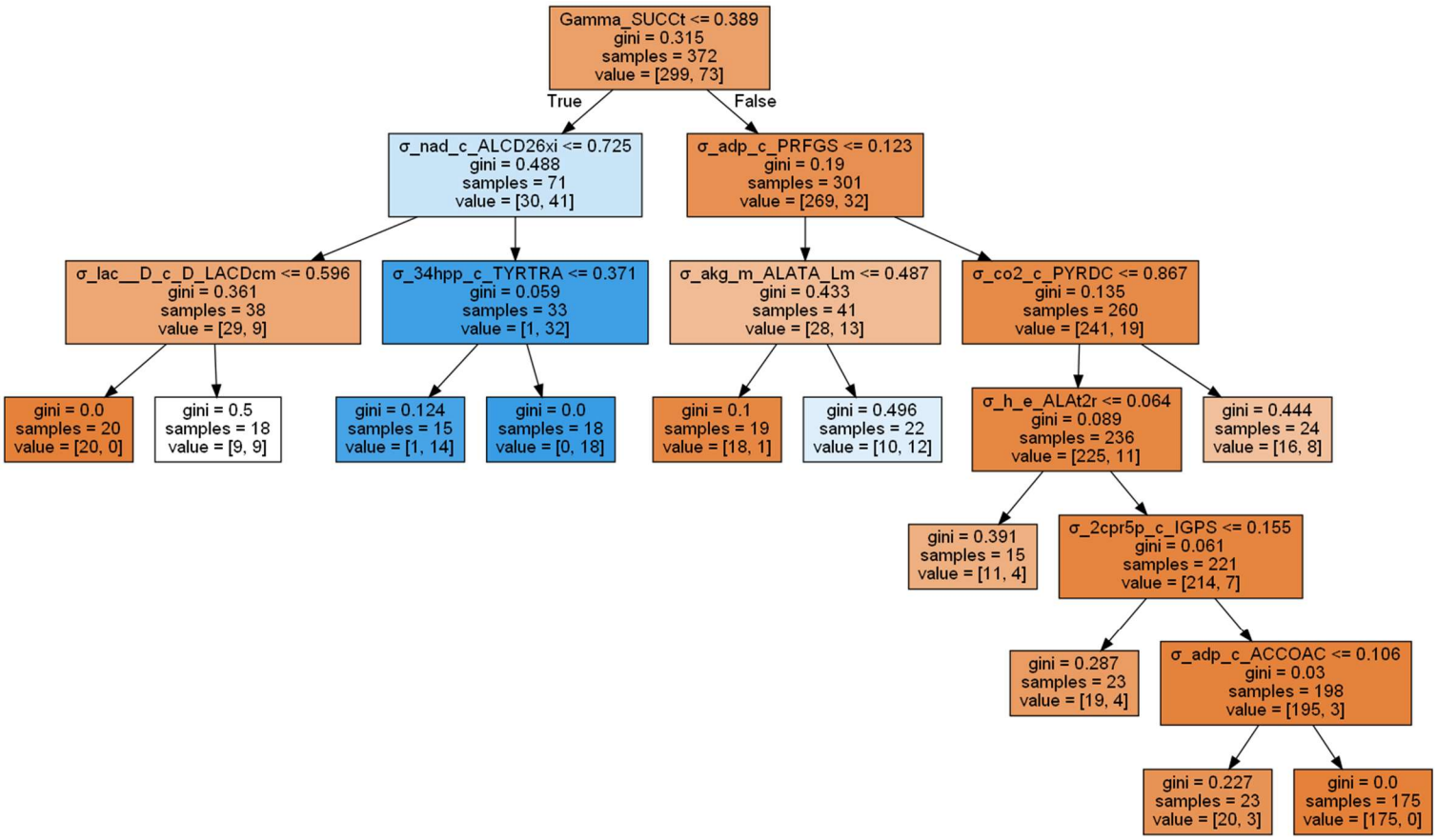

Figure 25:Decision tree for stable and unstable models with stability score 90%. Blue boxes correspond to stable model majority, orange boxes to unstable model majority and as we get closer to white we have equilibrium of models. Gini index illustrates to number

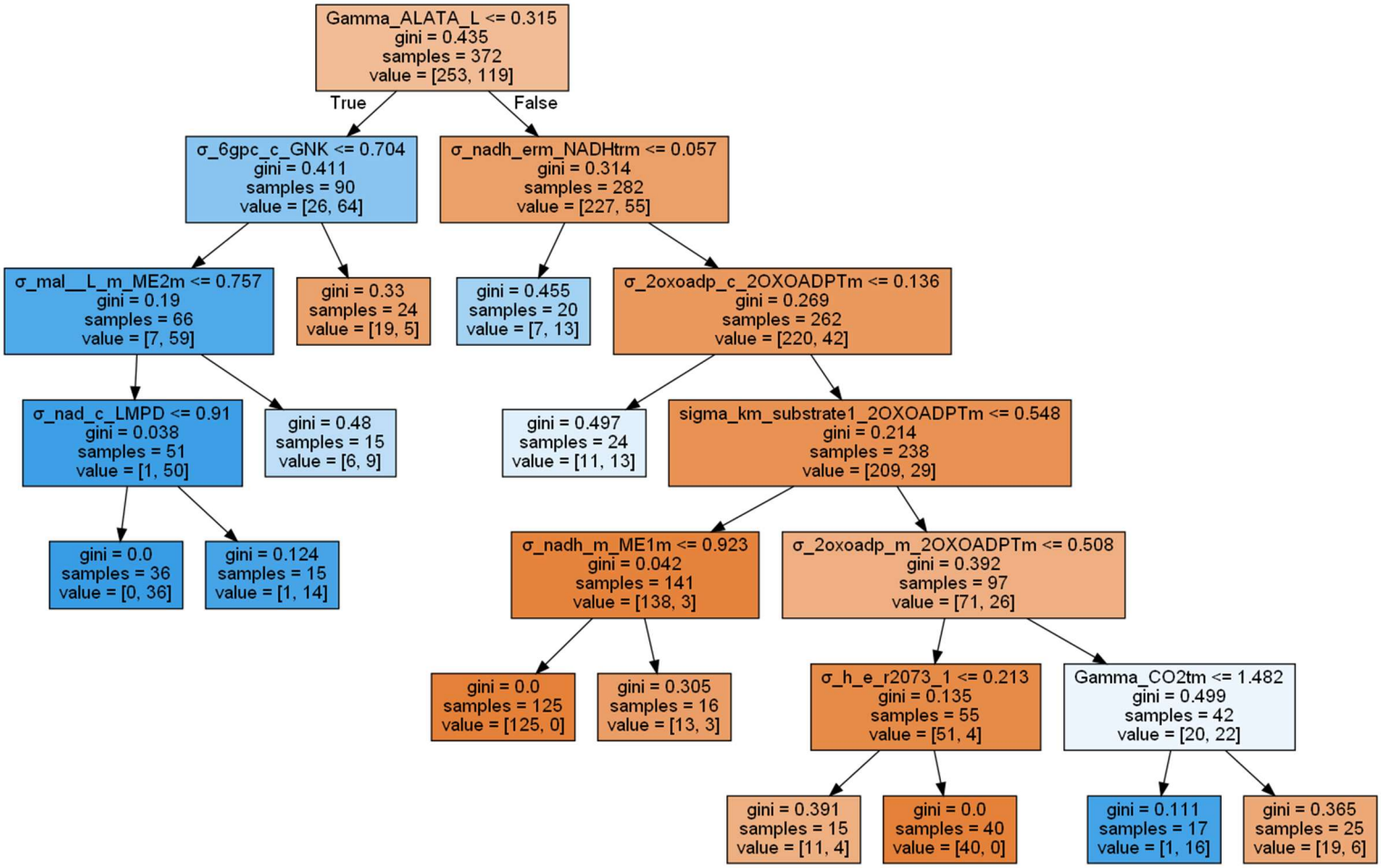

Figure 26:Decision Tree for stability score 80%. Blue boxes correspond to stable majority, orange to unstable majority and closer to white boxes correspond to almost same number of stable and unstable models

It is visible that for a different stability score the generated decision tree changes entirely. However, in both cases the first branch on the tree corresponds to a Gamma parameter. That is normal because the gamma value is calculated during the TFA sampling step. Having selected only 47 TFA samples to do the parameter inference leads to some TFA samples producing more stable kinetic models than others thus making the first branching a Gamma value.

For both stability cases the firsts branching classifies almost half the stable models with a very small percentage of unstable (42% and 28% for S=90% and S=80% accordingly). This signifies the importance of the steady state flux and concentration profiles prediction as they highly influence the stability of the generated models.

4.4 MCA

Having selected the kinetic models which are physiologically relevant and have showed great stability in the basins step we calculate the muconic acid flux control coefficients. To be more precise we calculated the flux control coefficients for the transport reaction of muconic acid from the cytosol department to the extracellular, cmm2tp. We could also use the reaction CaCatA (muconic acid production reaction) but it would yield the same results.

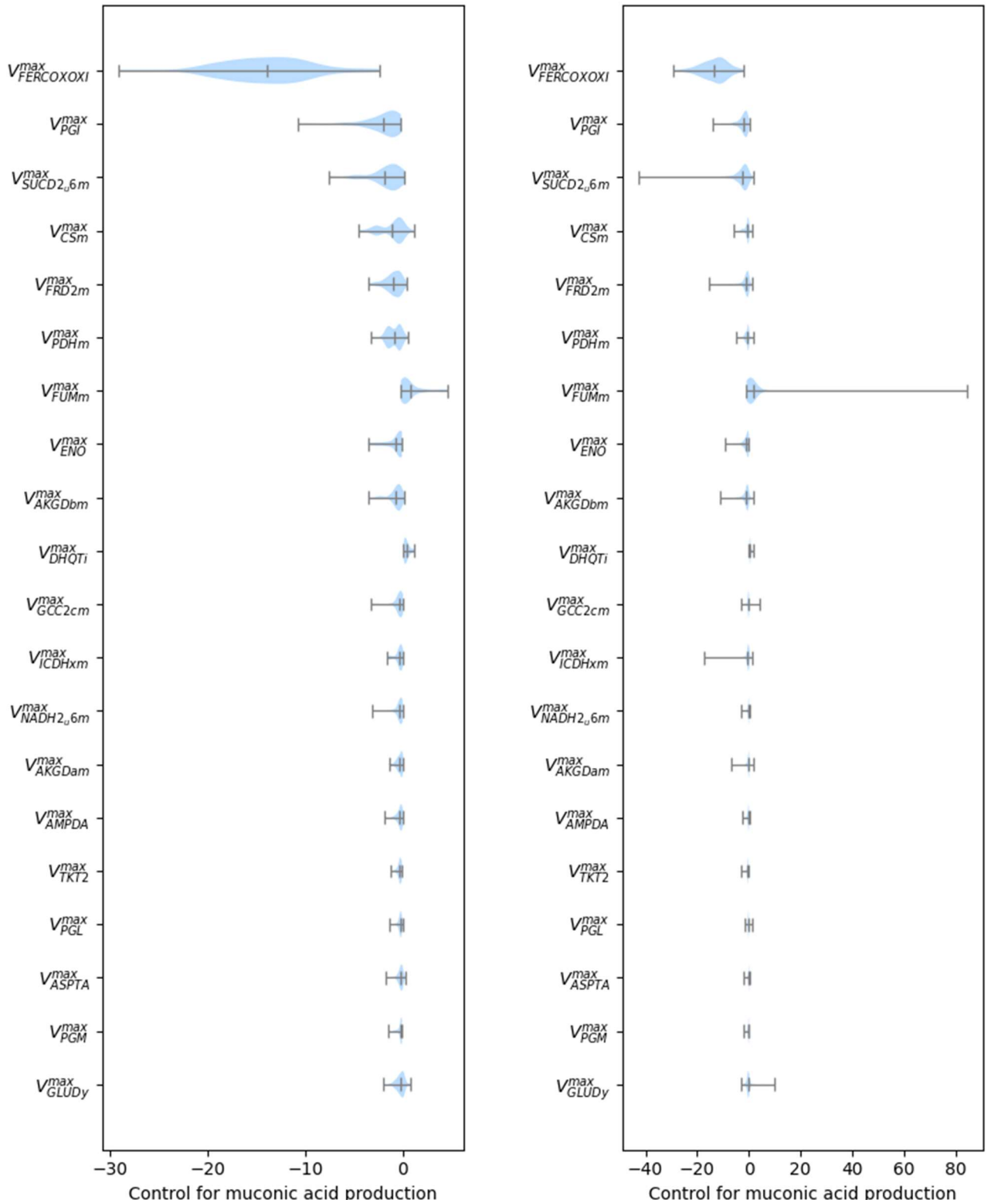

Figure 24: Violin plots for muconic acid production. The marker on the center represents the mean values and the two other points are the extreme values. The figure on the left represents the fccs of the 70 stable models and the figure on the right the fccs of the 366 models.

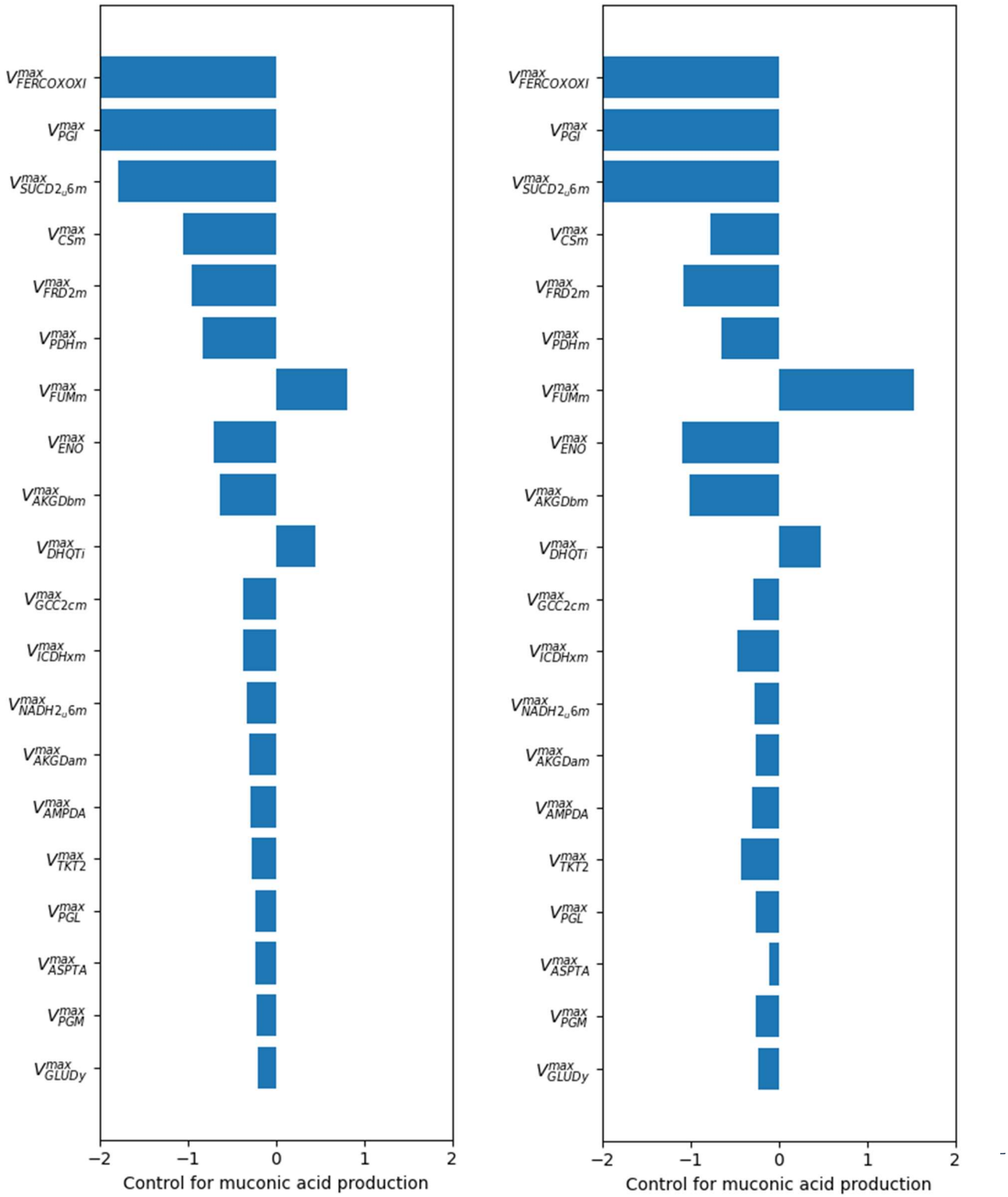

Figure 25: Bar plots of the mean values of fccs. On the left the fccs of the 70 stable models and on the right of the 366 models.

The large flux coefficients, for enzymes such as FERCOXOXI, SUCDu6m, PGI, FUMm, come as a result of small sampled fluxes.

Table 10: Information about the top 20 enzymes and the reactions they catalyze.

Enzyme	Subsystem	Reaction
FERCOXOXI	Oxidative phosphorylation	focytc_m + 1.266h_m + 0.25o2_m → ficytc_m + 0.5h2o_m + 0.633h_i
PGI	Glycolysis	g6p_c→ f6p_c
SUCD2u6m	Oxidative phosphorylation	q6_m + succ_m → fum_m + s_1535_m
FRD2m	Oxidative phosphorylation	nad_m + succ_m ← fum_m + h_m + nadh_m
PDHm	Oxidative phosphorylation	coa_m + nad_m + pyr_m → accoa_m + co2_m + nadh_m
CSm	Oxidative phosphorylation	accoa_m + h2o_m + oaa_m → cit_m + coa_m + h_m
FUMm	Oxidative phosphorylation	fum_m + h2o_m → mal__L_m
ENO	Glycolysis	2pg_c → h2o_c +pep_c
AKGDBm	Oxidative phosphorylation	coa_m + sdhlam_m → dhlam_m + succoa_m
DHQTi	Shikimate pathway	3dhq_c → 3dhsk_c + h2o_c
GCC2cm	Oxidative phosphorylation	dhlam_m + nad_m → h_m + lpam_m + nadh_m
ICDHxm	Oxidative phosphorylation	Icit_m + nad_m → akg_m +co2_m + nadh_m
NADH2u6m	Oxidative phosphorylation	h_m + nadh_m +q6_m → nad_m + s_1535_m
AKGDam	Oxidative phosphorylation	akg_m + h_m + lpam_m → co2_m + sdhlam_m
AMPDA	Purine and Pyrimidin biosynthesis	amp_c + h2o_c + h_c → imp_c + nh4_m
TKT2	Pentose Phosphate Pathway	e4p_c + xu5p__D_c → f6p_c + g3p_c
PGL	Pentose Phosphate Pathway	6pgl_c + h2o_c → 6pgc_c + h_c
ASPTA	Alanine and Aspartate Metabolism	akg_c + asp__L_c ← glu__L_c + oaa_c
PGM	Glycolysis	3pg_c → 2pg_c
GLUDy	Glutamate Metabolism	akg_c + h_c + nadph_c + nh4_C → glu__L_c +h2o_c +nadp_c

4.5 Enzyme Perturbations

To further quantify the effect on muconic acid flux of the enzymes with high flux control coefficients we performed enzyme perturbations. Based on the numeric sign of the fccs, we upregulated the enzyme if the sign was positive or downregulated if the sign was negative. The enzyme modification was simulated in the form a v_{max} value change, as the maximum rate is dependent on enzyme activity. For the new v_{max} value we calculated the new steady state muconic acid fluxes for all 70 stable models. It should be worth noting that for upregulation we set $v_{max,new} = 1.5v_{max,reference}$ and for downregulation $v_{max,new} = 0.5v_{max,reference}$.

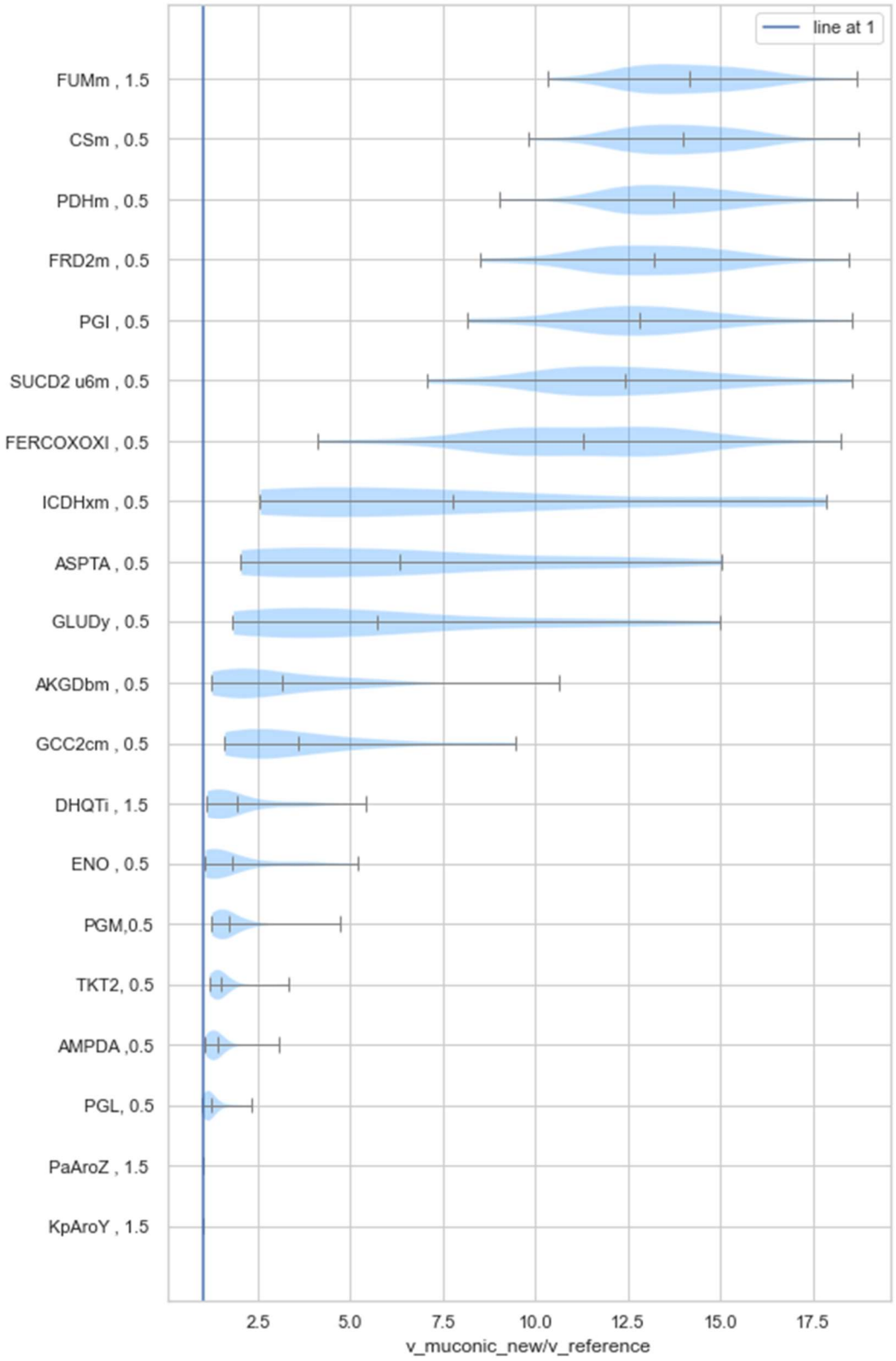

Figure 29: Violin plots of the effect of enzyme perturbations on muconic flux. Middle bar corresponds to the mean value. Both extremas also included.

Chapter 5. Conclusions and Future research

5.1 Developing large scale models

In the present study, we attempted to build large scale kinetic models for a muconic acid producing yeast. By bridging both approaches in strain design, the constraint-based methods (Thermodynamic Flux Balance Analysis) and the kinetic approach we managed through the ORACLE platform to generate physiologically relevant and stable kinetic models. These models were used to offer insight on the metabolic strategies that can be applied to increase muconic acid flux.

5.1.1 Genome scale curation and experimental data integration

The developed GEM corresponds to a mutant strain of S. Cerevisiae capable of producing muconic acid via shunting of the shikimic pathway. The model has thermodynamically curated and reduced. A kinetic model was created that incorporated all the information about the kinetic parameters and the kinetic expressions of the system's reactions. Experimental data such as growth rates, secretion rates, extracellular concentrations, were added to the GEM in the form of upper and lower bounds or reaction directionalities.

5.1.2 Kinetic models generation and pruning

In order to produce physiologically relevant kinetic models, we had to further constrain the metabolite concentration bounds to satisfy the pruning step criteria. During the TFA sampling step we identified 47 samples capable of producing relevant kinetic models and continued with generating 500 parameter sets for each TFA sample, thus generating a total of 23500 kinetic models. Out of those models, 366 were within physiological bounds and were used for further analysis.

5.1.3 Kinetic Models Stability

We continued with the stability test which consisted with 2 sets of 100 random perturbations each on two different ranges, a shorter and a wider one. As it is logical, models performed better in the shorter range as they had higher stability scores. Moreover, some kinetic models reached new steady states and by imposing a clustering algorithm we calculated the new centroids of the new steady states. An interesting observation is that in their majority kinetic models based on the same tfa sample reached the same alternative steady state.

With a stability score of 90%, 73 kinetic models were selected as candidates for the Metabolic Control Analysis and the enzyme perturbation step.

5.1.4 Decision Tree Analysis

Having picked out the kinetic models that showed robust cellular behavior we performed a decision tree analysis on the kinetic parameters of the generated models with the aim of indicating the parameters that seem to highly affect stability. For a stability score of 90% the most important kinetic parameter appears to be the thermodynamic displacement for SUCCt (Gamma<0.389) and the nad_c saturation for reaction ALCD26xi (sigma>0.725). If we wanted to increase the percentage of stable models during the generation step, we could employ those two bounds those two parameters. This iterative cycle would lead to more stable kinetic models and we would eventually constrain a plethora of kinetic parameters.

5.1.5 Metabolic Control Analysis

Using the 73 stable kinetic models we calculated the muconic flux control coefficients with the well-established MCA framework. We were able to identify the top 20 enzymes that affect muconic flux. Some attractive metabolic strategies that arisen were to increase flux to the Pentose Phosphate Pathway as to increase flux to e4p (PGI↓ ,TKT2↓,PGL↓) and to increase the dehydroshikimate flux (DHQTi↑). From the MCA results, we deducted that a bottleneck in muconic acid production is the deficiency of e4p in the cell. Moreover, we present the control coefficients for the all the physiological models and some differences are apparent such as the very larger control coefficients for FUMm and ENO.

5.1.6 Enzyme perturbations

For the enzymes that had large control coefficients for the muconic flux we performed some perturbations to calculate the effect on muconic flux. Although, some enzymes that are connected to the Electron Transport Chain (ETC) and it is infeasible to regulate we calculated the effect on muconic flux just for comparison. The candidate enzymes (PGI, TKT2, PGL, DHQTi) showed an increase in muconic flux with varying results. PGI downregulation by 0.5 resulted in 12.5 times bigger than reference muconic flux whereas TKT2 PGL downregulation by 0.5 resulted in 1.4 and 1.3 times accordingly bigger flux. DHQTi upregulation by 1.5 resulted in 1.8 times bigger than reference muconic flux. We also tested upregulating the heterologous enzymes (PaAroz,KpAroy) of the muconic pathway but the increase in muconic flux was insignificant.

5.2 Future Research

Kinetic modelling of such complex systems as cellular metabolism is a challenging task. However, in order to thin gap between lab observed strain yields and pilot scale yields we must be able to construct kinetic models which will be closer to reality. Although, constraint-based methods can offer a quick insight and great metabolic strategies for the increase of the desired flux they more often than ever cannot predict the dynamic aspect of the system. The formulation used in this thesis incorporated stoichiometry, kinetics and thermodynamics and produced models that are physiologically relevant and stable for a range of perturbations.

In order to produce more relevant kinetic models, we could set bounds to critical parameters that seem to affect model stability (e.g., for our case study gamma_SUCCt). The critical parameters can be classified using machine learning classification algorithms. Furthermore, the experimental data for the growth rate, secretion rates could be used to further screen out the generated models. Pattern recognition algorithms could be employed to identify kinetic models that seem to have the same behavior as the experimentally observed trajectory.

References

1. d'Espaux, L. et al. Engineering high-level production of fatty alcohols by Saccharomyces cerevisiae from lignocellulosic feedstocks. Metab. Eng. 42, 115–125 (2017).

2. Kokossis, A. C., Tsakalova, M. & Pyrgakis, K. Design of integrated biorefineries. Comput. Chem. Eng. 81, 40–56 (2015)

3. Home - 2019 - United Nations Sustainable Development. Available at: https://www.un.org/sustainabledevelopment/. (Accessed: 21st September 2019)

4. Lokko, Y. et al. Biotechnology and the bioeconomy—Towards inclusive and sustainable industrial development. N. Biotechnol. 40, 5–10 (2018).

5. Gustavsson, M. & Lee, S. Y. Prospects of microbial cell factories developed through systems metabolic engineering. Microb. Biotechnol. 9, 610–617 (2016).

6. Caspeta, L. & Nielsen, J. Economic and environmental impacts of microbial biodiesel. Nat. Biotechnol. 2013 319 (2013).

7. Lee, J. W. et al. Systems metabolic engineering of microorganisms for natural and non-natural chemicals. Nat. Chem. Biol. 8, 536–546 (2012).

8. Long, M. R., Ong, W. K. & Reed, J. L. Computational methods in metabolic engineering for strain design. Curr. Opin. Biotechnol. 34, 135–141 (2015).

9. Maranas, C. D. & Zomorrodi, A. R. Optimization methods in metabolic networks.

10. Burgard, A. P., Pharkya, P. & Maranas, C. D. Optknock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotechnol. Bioeng. 84, 647–657 (2003).

11. Wuest, D., Hou, S., & Lee, K. (2011). Metabolic Engineering. Comprehensive Biotechnology.

12. Lee, J. W., Na, D., Park, J. M., Lee, J., Choi, S., & Lee, S. Y. (2012). Systems metabolic engineering of microorganisms for natural and non-natural chemicals. *Nature Chemical Biology*.

13. Xie, N. Z., Liang, H., Huang, R. B., & Xu, P. (2014). Biotechnological production of muconic acid: current status and future prospects. *Biotechnology Advances*,

14. Eudes, A., Berthomieu, R., Hao, Z., Zhao, N., Benites, V. T., Baidoo, E. E., & Loqué, D. (2018). Production of muconic acid in plants. *Metabolic Engineering*, *46*, 13–19.

15. Khalil, I., Quintens, G., Junkers, T., & Dusselier, M. (2020). Muconic acid isomers as platform chemicals and monomers in the biobased economy. *Green Chemistry*, *22*(5), 1517–1541.

16. Harwood, C. S., & Parales, R. E. (1996). THE β-KETOADIPATE PATHWAY AND THE BIOLOGY OF SELF-IDENTITY. *Annual Review of Microbiology*, *50*(1), 553–590.

17. van Duuren, J. B. J. H., Wijte, D., Karge, B., Martins Dos Santos, V. A., Yang, Y., Mars, A. E., & Eggink, G. (2011). pH-stat fed-batch process to enhance the production of cis, cis-

muconate from benzoate by Pseudomonas putida KT2440-JD1. *Biotechnology Progress*, *28*(1), 85–92.

18. Bang, S. G., & Choi, C. Y. (1995). DO-stat fed-batch production of cis, cis-muconic acid from benzoic acid by Pseudomonas putida BM014. *Journal of Fermentation and Bioengineering*, *79*(4), 381–383.

19. Kaneko, A., Ishii, Y., & Kirimura, K. (2011). High-yield Production ofcis,cis-Muconic Acid from Catechol in Aqueous Solution by Biocatalyst. *Chemistry Letters*, *40*(4), 381–383.

20. Joffres, B., Laurenti, D., Charon, N., Daudin, A., Quignard, A., & Geantet, C. (2013). Thermochemical Conversion of Lignin for Fuels and Chemicals: A Review. *Oil & Gas Science and Technology – Revue d'IFP Energies Nouvelles*, *68*(4), 753–763.

21. Bui V, Lau MK, MacRae D, Schweitzer D. Methods for producing isomers of muconic acid and muconate salts. US Patent 20130030215 A1, 2013.

22. Sun, X., Lin, Y., Huang, Q., Yuan, Q., & Yan, Y. (2013). A Novel Muconic Acid Biosynthesis Approach by Shunting Tryptophan Biosynthesis via Anthranilate. *Applied and Environmental Microbiology*, *79*(13), 4024–4030.

23. Bongaerts, J., Krämer, M., Müller, U., Raeven, L., & Wubbolts, M. (2001). Metabolic Engineering for Microbial Production of Aromatic Amino Acids and Derived Compounds. *Metabolic Engineering*, *3*(4), 289–300.

24. Ghosh, S., Chisti, Y., & Banerjee, U. C. (2012). Production of shikimic acid. *Biotechnology Advances*, *30*(6), 1425–1431.

25. Nielsen, J. Yeast Systems Biology: Model Organism and Cell Factory. Biotechnol. J. 14, 1800421 (2019).

26. Mell, J. C., & Burgess, S. M. (2003). Yeast as a Model Genetic Organism. *Encyclopedia of Life Sciences*.

27. Mahadevan, R. & Schilling, C. H. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metab. Eng. 5, 264–276 (2003)

28. Pinzon, W., Vega, H., Gonzalez, J., & Pinzon, A. (2018). Mathematical Framework Behind the Reconstruction and Analysis of Genome Scale Metabolic Models. *Archives of Computational Methods in Engineering*, *26*(5), 1593–1606.

29. Oberhardt, M. A., Palsson, B., & Papin, J. A. (2009). Applications of genome-scale metabolic reconstructions. *Molecular Systems Biology*, *5*(1), 320.

30. Lewis, N. E., Nagarajan, H. & Palsson, B. O. Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. Nat. Rev. Microbiol. 10, 291–305 (2012).

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

32. Soh, K. C. & Hatzimanikatis, V. Constraining the Flux Space Using Thermodynamics and Integration of Metabolomics Data. in 49–63 (Humana Press, New York, NY, 2014).

33. Henry CS, Broadbelt LJ, Hatzimanikatis V. Thermodynamics-based metabolic flux analysis. Biophys J.

34. Henry CS, Jankowski MD, Broadbelt LJ, Hatzimanikatis V. Genome-scale thermodynamic analysis of *Escherichia coli* metabolism. Biophys J.

35. Chakrabarti, A., Miskovic, L., Soh, K. C., & Hatzimanikatis, V. (2013). Towards kinetic modeling of genome-scale metabolic networks without sacrificing stoichiometric, thermodynamic and physiological constraints. *Biotechnology Journal*, *8*(9), 1043–1057.

36. Mendes, P., Stanford, N. J., & Smallbone, K. (2011). Kinetic modelling of large-scale metabolic networks. *Proceedings of the 9th International Conference on Computational Methods in Systems Biology - CMSB '11*. Published.

37. Teusink B, Passarge J, Reijenga CA, et al. Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry. *Eur J Biochem*.

38. Visser D, Heijnen JJ. Dynamic simulation and metabolic re-design of a branched pathway using linlog kinetics. Metab Eng 2003;5:164–176.

39. Smallbone K, Simeonidis E, Broomhead DS, Kell DB. Something from nothing: Bridging the gap between constraint-based and kinetic modelling.

40. Liebermeister W, Klipp E. Bringing metabolic networks to life: Convenience rate law and thermodynamic constraints. Theor Biol Med Model 2006;3:1–11.

41. H, Burns JAE. The control of flux. Symp Soc Exp Biol 1973;27:65–104.

42. Liebermeister W, Uhlendorf J, Klipp E. Modular rate laws for enzymatic reactions: Thermodynamics, elasticities and Implementation. Bioinformatics 2010;26:1528–1534.

43. Lubitz T, Schulz M, Klipp E, Liebermeister W. Parameter balancing in kinetic models of cell metabolism. J Phys Chem B 2010;114:16298–16303.

44. Andreozzi, S., Miskovic, L., & Hatzimanikatis, V. (2016). iSCHRUNK – In Silico Approach to Characterization and Reduction of Uncertainty in the Kinetic Models of Genome-scale Metabolic Networks. *Metabolic Engineering*, *33*, 158–168.

45. Foster, C. J., Wang, L., Dinh, H. V., Suthers, P. F., & Maranas, C. D. (2021). Building kinetic models for metabolic engineering. *Current Opinion in Biotechnology*, *67*, 35–41.

46. Stanford, N. J., Lubitz, T., Smallbone, K., Klipp, E., Mendes, P., & Liebermeister, W. (2013). Systematic Construction of Kinetic Models from Genome-Scale Metabolic Networks. *PLoS ONE*, *8*(11), e79195.

47. Tokic, M., Hatzimanikatis, V., & Miskovic, L. (2020). Large-scale kinetic metabolic models of Pseudomonas putida KT2440 for consistent design of metabolic engineering strategies. *Biotechnology for Biofuels*, *13*(1).

48. Miskovic, L., & Hatzimanikatis, V. (2010). Production of biofuels and biochemicals: in need of an ORACLE. *Trends in Biotechnology*, *28*(8), 391–397.

49. Wang, L., Birol, I., & Hatzimanikatis, V. (2004). Metabolic Control Analysis under Uncertainty: Framework Development and Case Studies. *Biophysical Journal*, *87*(6), 3750–3763.

50. Ataman, M., Hernandez Gardiol, D. F., Fengos, G., & Hatzimanikatis, V. (2017). redGEM: Systematic reduction and analysis of genome-scale metabolic reconstructions for development of consistent core metabolic models. *PLOS Computational Biology*, *13*(7), e1005444.

51. Ataman, M., & Hatzimanikatis, V. (2017). lumpGEM: Systematic generation of subnetworks and elementally balanced lumped reactions for the biosynthesis of target metabolites. *PLOS Computational Biology*, *13*(7), e1005513.

52. Wang, G., ØZmerih, S., Guerreiro, R., Meireles, A. C., Carolas, A., Milne, N., Jensen, M. K., Ferreira, B. S., & Borodina, I. (2020). Improvement of cis,cis-Muconic Acid Production in Saccharomyces cerevisiae through Biosensor-Aided Genome Engineering. *ACS Synthetic Biology*, *9*(3), 634–646.

53. Herrmann, H. A., Dyson, B. C., Vass, L., Johnson, G. N. & Schwartz, J.-M. Flux sampling is a powerful tool to study metabolism under changing environmental conditions. npj Syst. Biol. Appl. 5, 32 (2019).

54. Saa, P. A. & Nielsen, L. K. ll-ACHRB: a scalable algorithm for sampling the feasible solution space of metabolic networks. Bioinformatics 32, 2330–2337 (2016).

55. Schellenberger, J. & Palsson, B. Ø. Use of Randomized Sampling for Analysis of Metabolic Networks. J. Biol. Chem. 284, 5457–5461 (2009).

56. Miskovic, L., Hatzimanikatis, V., Modeling of uncertainties in biochemical reactions. Biotechnol. Bioeng. 2011, 108, 413–423.

57. Heinrich, R., Schuster, S., The Regulation of Cellular Systems, Chapman & Hall, New York 1996.

58. Hatzimanikatis, V., Floudas, C. A., Bailey, J. E., Analysis and design of metabolic reaction networks via mixed-integer linear optimization. AIChE J. 1996, 42, 1277–1292.

59. Hatzimanikatis, V., Bailey, J. E., MCA has more to say. J. Theor. Biol. 1996, 182, 233–242.

60. Hatzimanikatis, V., Bailey, J. E., Effects of spatiotemporal variations on metabolic control: Approximate analysis using (log)linear kinetic models. Biotechnol. Bioeng. 1997, 54, 91–104.

61. Hartigan, J. A. & Wong, M. A. Algorithm AS 136: A K-Means Clustering Algorithm. Appl. Stat. 28, 100 (1979).

62. Hatzimanikatis, V., C. A. Floudas, and J. E. Bailey. 1996. Analysis and design of metabolic reaction networks via mixed-integer linear optimization. AICHE J. 42:1277–1292.

63. Reder, C. 1988. Metabolic control theory: a structural approach. J. Theor. Biol. 135:175–201

64. Lu, H., Li, F., Sánchez, B. J., Zhu, Z., Li, G., Domenzain, I., Marcišauskas, S., Anton, P. M., Lappa, D., Lieven, C., Beber, M. E., Sonnenschein, N., Kerkhoven, E. J., & Nielsen, J. (2020). Author Correction: A consensus S. cerevisiae metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism. *Nature Communications*, *11*(1).