Navigating and Managing the Complexity of Genome Scale Metabolic Networks for Studies in Cellular Physiology and Industrial Biotechnology

THÈSE N^{O} 7102 (2016)

PRÉSENTÉE LE 28 OCTOBRE 2016

À LA FACULTÉ DES SCIENCES DE BASE

LABORATOIRE DE BIOTECHNOLOGIE COMPUTATIONNELLE DES SYSTÈMES

PROGRAMME DOCTORAL EN CHIMIE ET GÉNIE CHIMIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Meriç ATAMAN

acceptée sur proposition du jury:

Prof. M. Dal Peraro, président du jury
Prof. V. Hatzimanikatis, directeur de thèse
Prof. N. Price, rapporteur
Prof. R. Mahadevan, rapporteur
Prof. U. Sauer, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2016

To Nariye…

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my supervisor Vassily Hatzimanikatis for his support of my Ph.D. study and research, for his patience, motivation, enthusiasm, and immense knowledge. Apart from being the best advisor you can imagine, he is also there whenever you need any advice on non-scientific things, for deep discussions about history, politics, literature and so on. I have learnt a lot from him, and spent the best years of my education life in his lab. Thank you very very much Vassily!

I would also like to thank Professor Costas Panayiotou for his supervisory role while he was visiting us in LCSB and for the excellent collaboration we had at the time.

My acknowledgement also goes to Christine Kupper and Anne Lene Odegaard for the administrative support and their help to get organized in so many occasions.

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Krishna Mahadevan, Prof. Nathan Price, and Prof. Uwe Sauer, for their encouragement, insightful comments, and questions.

I would like to thank the present and the past members of the LCSB group: Firstly, I would like to say a very big thank you to Dr. Ljubisa Miskovic, not only for his continuous support and help on scientific topics, but also for being always positive and cheerful and of course for drinking sessions. My industrial project partner, who also contributed to this thesis, Noushin Hadadi, I would like to thank her for being a very good collaborator and friend, and reminding me all the time the things I forgot to do!

And my dear lab/conference mate Georgios Fengos, thank you very much for all the small (or big?) talks we had, and for being positive all the time! Tiziano, you know I am in Bar Tabac, so if you want to see me, just stop by.

Daniel, Stepan, Jasmin, Anush, Milenko, Tuure, Georgios S., Vikash, thank you very much guys for all the good times we spent together and creating a wonderful environment in the lab! I would like to thank also Keng, for helping me to learn the things he did, and being always friendly.

Special thanks go to my best friends in Lausanne, Emre, Basak and Ismail, with whom I

spent amazing time in here. And the Turkish gang! Kerem, Can, Bugra, Berker, Cem and all, thank you very much guys! Let's have some beer.

I think I cannot thank enough to my parents, my mom and dad, the only thing I can say is: Hersey icin cok tesekkurler, Trabzon'un kucucuk bir koyunden bizi buralara kadar getirdiginiz icin. Hayatimiza sinirlar koymadiginiz icin. Boyle guzel insanlar oldugunuz icin.

My brother, Caglar. He is kind of a guy that will always open your horizon, but with the friendliest way you can think of. Thank you very much for all, for music, for discussions, for the food and for being such a great brother, more importantly a great friend. And, don't serve the best whiskeys to anyone else if I am not there, please.

And my wife, Nariye… I don't know if I will be able to describe how deeply thankful I am to you. I couldn't do anything I achieved till now without you. You changed the laziest person on earth, to someone with a Ph.D. Whenever I felt down, you were there to encourage me and put back the smile on my face. Thank you very much for making my life full of happiness and wonderful. I cannot think of a life without you. My love, my best friend.

23.06.2016, Lausanne

ABSTRACT

Metabolism is a network of biochemical reactions that converts carbon sources to cellular fuels and building blocks. Metabolism of many organisms has been studied for many decades. These studies define metabolism as a very large and complex network. Stoichiometric models have been used extensively to study these complex networks. With the introduction of Genome Scale Models (GEMs), the studies on metabolism entered a new era, and gained a more systems approach. GEMs have been used in a broad range of areas, from cancer research to industrial applications. By using GEMs, it is possible to study metabolism with an input/output manner, and decompose it to individual fluxes by detailed stoichiometric definition. However, due to the large degrees of freedom and the underdetermined nature of GEMs, it is crucial to develop methods to further constrain these complex networks to reveal the actual state of the metabolism. Integrating thermodynamics constraints to metabolic networks is a popular and powerful approach to address this need. Moreover, these methods allow the incorporation of steady-state metabolite concentrations into GEMs, which cannot be achieved by other methods such as Flux Balance Analysis (FBA). In this thesis, we firstly discussed different methods to incorporate this constraint into metabolic models and used the most comprehensive one, Thermodynamics-based Flux Analysis (TFA) for studies in the next chapters. Firstly, we used TFA to study the overall behavior of *E. coli* in terms of bioenergetics efficiency, P:O ratio and Gibbs free energy dissipation and revealed their connection. Following this analysis, we focused on the complexity that emerges from the size of the GEMs. GEMs are composed of large metabolic modules, called subsystems or pathways. In this thesis, we aimed to re-define the pathway definition by generating subnetworks for the synthesis of biomass building blocks using lumpGEM, a tool that extracts parts of metabolism for certain tasks, such as synthesis of an amino acid. lumpGEM identified additional reactions from different parts of metabolism along with textbook pathways for synthesis of many biomass building blocks. lumpGEM also builds lumped reactions for the gener-

ated subnetworks, which represent the subnetwork with 1 overall reaction, thus reducing complexity. redGEM uses this property to build reduced models, moreover it redefines the central carbon metabolism definition, and build core models consistent with their GEMs. These reduced models are valuable platforms for many studies, such as kinetic modeling, FBA/TFA studies and comparison of central carbon networks among different organisms. We used a reduced model of *E. coli* to study further the characteristics of the core models, and their still big solution space by enumerating all possible Flux Directionality Profiles (FDP). We identified the effect of directionality of reactions on overall network behavior, such as specific growth rate. We finally focused on the metabolic capabilities of *E. coli* and identified possible biotransformation that *E. coli* can perform by using BNICE.ch. We built a *super* network for *E. coli* and studied its characteristics for biomass production and metabolic gaps. As a conclusion, in the last chapter we discuss the potential applications of the methods and tools that we developed in this thesis.

Keywords: metabolism, metabolic networks, complexity, thermodynamics, energy dissipation, subsystems/pathways, reduction, directionality, central carbon metabolism

RÉSUMÉ

Le métabolisme est un réseau de réactions biochimiques qui convertit les sources de carbone en combustibles cellulaires et composantes. Le métabolisme de nombreux organismes a été étudié pendant de nombreuses décennies. Ces études définissent le métabolisme comme un réseau très vaste et complexe. Des modèles stœchiométriques ont été abondamment utilisés pour étudier ces réseaux complexes. Avec l'introduction de modèles à l'échelle des génomes (GEM), les études sur le métabolisme sont entrés dans une nouvelle ère, et ont bénéficié d'une approche plus systémique. Les GEM ont été utilisés dans un large éventail de domaines, de la recherche sur le cancer à des applications industrielles. En utilisant des GEM, il est possible d'étudier le métabolisme d'une manière entrées/sorties, et de le décomposer à des flux individuels par définition stœchiométrique détaillée. Cependant, en raison des grands degrés de liberté et de la nature sous-déterminée des GEM, il est crucial de développer des méthodes pour contraindre davantage ces réseaux complexes afin de révéler l'état réel du métabolisme. L'intégration de contraintes thermodynamiques aux réseaux métaboliques est une approche répandue et puissante pour répondre à ce besoin. De plus, ces méthodes permettent l'incorporation de concentrations de métabolites à l'état stationnaire dans les GEM, qui ne peuvent pas être atteints par d'autres méthodes telles que l'analyse de l'équilibre des flux (FBA). Dans cette thèse, nous avons tout d'abord discuté de différentes méthodes pour intégrer cette contrainte dans les modèles métaboliques et utilisé la plus complète, l'analyse des flux basée sur la thermodynamique (TFA) pour les études dans les chapitres suivants. Tout d'abord, nous avons utilisé TFA pour étudier le comportement global de E. coli en termes d'efficacité de la bioénergétique, rapport P:O et la dissipation de l'énergie libre de Gibbs et révélé leur connexion. Suite à cette analyse, nous nous sommes concentrés sur la complexité qui se dégage de la taille des GEM. Les GEM sont composés de grands modules métaboliques, appelés sous-systèmes ou voies. Dans cette thèse, nous avons cherché à redéfinir la définition de la voie en créant des sous-réseaux pour la synthèse de composantes de la biomasse en utilisant lumpGEM, un outil qui extrait les parties du métabolisme pour certaines tâches, telles que la synthèse d'un acide aminé. lumpGEM a identifié des réactions supplémentaires provenant de différentes

parties du métabolisme ainsi que les voies typiques pour la synthèse de nombreuses composantes de la biomasse. lumpGEM construit aussi des réactions regroupées pour les sous-réseaux générés, qui représentent le sous-réseau avec une seule réaction globale, réduisant ainsi la complexité. redGEM utilise cette propriété pour construire des modèles réduits, mais aussi pour revoir la définition du métabolisme central du carbone, et pour construire des modèles de base compatibles avec leurs GEMs. Ces modèles réduits sont des plates-formes utiles pour de nombreuses études, telles que la modélisation cinétique, des études FBA/TFA et la comparaison des réseaux centraux de carbone entre les différents organismes. Nous avons utilisé un modèle réduit de E. coli pour étudier davantage les caractéristiques des modèles de base, et leur grand espace de solution en dénombrant tous les profils de direction des flux (FDP) possibles. Nous avons identifié l'effet de la directivité des réactions sur le comportement global du réseau, tels que les taux de croissance. Nous avons finalement mis l'accent sur les capacités métaboliques de E. coli et identifié la biotransformation possible que E. coli peut effectuer à l'aide de BNICE.ch. Nous avons construit un super réseau pour E. coli et étudié ses caractéristiques pour la production de la biomasse et des lacunes métaboliques. En conclusion, dans le dernier chapitre, nous discutons des applications potentielles des méthodes et des outils que nous avons développés dans cette thèse.

Mots-clés: métabolisme, réseaux métaboliques, complexité, Thermodynamique, dissipation d'énergie, réduction, directivité, métabolisme central carboné

TABLE OF CONTENTS

LIST OF TABLES

Appendix Tables

LIST OF FIGURES

Abbreviations

S: Stoichiometric matrix

v: Flux vector

v_{min}: Vector containing lower bounds for fluxes

v_{max}: Vector containing upper bounds for fluxes

γ: Coefficient for enforcing P:O Ratio

ΔG^0_{rxn}: Standard Gibbs free energy of reaction

$\Delta G'^0_{rxn}$: Standard transformed Gibbs free energy of reaction

$\Delta G'_{rxn}$: Gibbs free energy of reaction

$\ln(c_i)$: Logarithmic concentration of metabolite *i*

ϕ: Gibbs free energy dissipation

μ_{max}: Maximum specific growth rate

S_{min}: Minimum subnetwork to produce a target metabolite from core network

n_i: Stoichiometric coefficient of metabolite *i*

D: Degree of connection *D* to connect subsystems

FU^j_{rxn}: Forward binary use variable for reaction *j*

BU^j_{rxn}: Backward binary use variable for reaction *j*

C: large arbitrary constant to enforce lower flux bound constraint

R: Molar gas constant

T: Temperature

Chapter 1 - Introduction

Metabolism is the set of biochemical transformations that sustain the life of organisms. Metabolism supplies the fuel for cellular processes; it supplies the precursors for the synthesis of complex molecules such as proteins, lipids, nucleic acids, and some carbohydrates. Metabolism and metabolic pathways have been studied in details for decades and the underlying mechanisms and biotransformations for many organisms have been revealed. These studies focused on a wide spectrum of areas from clinical research to industrial large-scale production. Metabolic disorders have been shown to be connected to many diseases, such as cancer, obesity, diabetes, etc. [1] Metabolic interactions in the gut microbiota have emerged as a promising research area in the last decade to understand the role of these organisms in the host human[2] and to reveal their connection with diseases[3]. Moreover, metabolism is an excellent platform for *white biotechnology*[4]*;* a research area that has close connections with industry for the replacement of petroleum based chemical synthesis with bio-based synthesis methods. The focus of scientific studies on biology has shifted to many different areas in years, metabolism yet kept its importance and proved to be a promising and productive research area.

The accumulated knowledge through decades indicates the complexity of the metabolism. Metabolic genes constitute a significant portion of the genes encoded by the organisms. *S. cerevisiae* has ~6600 genes, and the consensus Yeast 6 metabolic network[5] includes 900 genes. *E. coli* has ~4400 genes, and the latest metabolic reconstruction, iJO1366[6] includes 1366 genes, which constitutes 32% of all genes. Gene to Protein to Reactions (GPRs) associations, which form the bridge between the genome and biochemistry, reveals thousands of metabolic reactions, defining a very complex network. Complexity of metabolism does not only emerge from the *number* of metabolic reac-

tions, but from the tight connections of metabolites that participate in these reactions. Well-characterized metabolic networks are nested[7], which means that the metabolic networks are not only composed of linear pathways that supply the precursors to the next pathway. A metabolite can appear in different parts of the metabolism, being used by several different enzymes as substrate or product. Although the pathways reported in databases for certain functions, such as biosynthesis of amino acids, lipids, nucleoside-3-phosphates are linear, it does not necessarily mean that the enzymes that are active for the synthesis of this particular metabolite along those pathways are the only ones that are reported in those linear representations.

To study this complexity computationally, metabolism is represented through metabolic networks, and metabolic networks are built through GPR associations[8]. However, the list of reactions obtained through this process is not enough to study the metabolism computationally. The solution for this need, stoichiometric models has been very popular tools to study these complex networks through mathematical representations. These models are built through a stoichiometric matrix, which keeps the information of metabolites and their stoichiometric participation in every reaction in the network, and in early examples, they represented a certain part of the metabolism picked with an *ad hoc* manner. These stoichiometric models evolved into Genome Scale Metabolic Reconstructions (GEMs), which encapsulate all metabolic capabilities obtained from GPR associations with a bottom-up approach[9]. These GEMs are accompanied with a composition of 1 gDW of cell on metabolite basis (amino acids, lipids, etc.)[10], and the boundary conditions, i.e. media composition and possible secreted metabolites. This approach enabled researchers to study on the overall behaviour of metabolic networks in an input/output manner such as biomass growth or production of specific biochemicals. Moreover, it gave the opportunity to study the individual contribution of the metabolic fluxes that constitute these metabolic outputs.

Even though metabolism is nested and metabolic pathways are not linear, the relation between metabolites and reactions represented by stoichiometric matrix is a set of linear equations. This linearity, along with the linear representation of the biomass composition brings a mathematical advantage for simulations for optimum biomass produc-

tion, target metabolite biosynthesis, or any other cellular objectives. Along with the quasi-steady state assumption, which assumes that there is no accumulation of any metabolite in the system, metabolic networks are mathematically represented by the following formulation:

$$\boldsymbol{S}.\boldsymbol{v} = \mathbf{0}$$

$$\boldsymbol{v_{i,min}} < \boldsymbol{v_i} < \boldsymbol{v_{i,max}}$$

where S is the stoichiometric matrix; v is the flux vector that defines the flux through every reaction defined in the metabolic network. The right hand side of the equation states that there is no accumulation of metabolites in the network. v_{min} and v_{max} represent the allowable ranges for fluxes for these metabolic reactions.

A well-connected metabolic network is underdetermined, which means that the number of equations (metabolite mass balances) is bigger than the number of variables (reactions)[11]. This brings a complexity to computational studies, since underdetermined systems has infinite number of solutions, even though in linear space they have only 1 global optimum for a given objective. In biological terms, this means that there are infinite number of flux profiles that can satisfy the same cellular objective.

Researchers have been working on different methods to understand and characterize the complexity, and formulated different approaches to reduce it. The very first approach is to identify the reaction directionalities in the metabolic networks. Most of the enzymes can catalyze a certain reaction in both direction, and there are a few enzymes characterized as kinetically irreversible. To reduce the complexity that emerges from bidirectionality and to recude the solution space, one of the most effective methods is to integrate thermodynamics into metabolic networks. Thermodynamics, through Gibbs Free Energy of reactions, has the capability to reveal if a certain enzyme can catalyze a reaction in both directions given the cellular conditions, such as pH, ionic strength, etc. [12]. This analysis evolved into network thermodynamics for metabolic networks,

which globally reduce the solution space with energy constraints[13-15]. In Chapter 2, we have reviewed different approaches to integrate thermodynamics into metabolic networks, and we discussed in details our approach, and its advantages over other methods developed in the last 10 years along with other applications and future perspective.

Thermodynamics, up to now, is introduced to the analysis of metabolic networks to suggest the directionality for the reactions and by network thermodynamics to reduce the solution space. Moreover, it has been used to compute thermodynamically feasible concentration ranges for metabolites, thus distance from thermodynamic equilibrium[16]. However, it has not been used systematically to understand the overall behaviour of the cell in metabolic networks, such as bioenergetics efficiency. It has been long known that the Gibbs free energy dissipation can be used to predict the cellular growth[17], and the relation between these two properties has been shown experimentally and computationally through different type of modeling. However, such studies have not been carried out in the context of stoichiometric modeling and Flux Balance analysis. There are a few examples, where the bioenergetics efficient concept is used to metabolic modeling through P:O Ratio, which is the number of ATP molecules produced per number of oxygen atoms uptaken by the cell[18]. In Chapter 3, we have expanded the scope of stoichiometric modeling and introduced the concept of free energy dissipation through TFA and its relation with P:O Ratio. We have determined the relation between those two bioenergetics concepts and revealed how they can explain physiology of optimal growth and by-product formation of *E. coli*. In addition, we have determined the enzymes that are mainly contributing to the free energy dissipation for the cell.

Thermodynamics is a powerful tool to reduce the solution spaces for metabolic networks; however, it does not have any effect on the network topology. Metabolic networks are usually defined as the sum of modules[19] or pathways[20], such as amino acid biosynthesis, and subsystems, glycolysis, TCA cycle etc. However, these modules or subsystems are not isolated from each other and share many metabolites, even reactions, which causes a different form of complexity. To manage this, in Chapter 4, we focused on cellular growth with a novel approach that aimed to re-define the pathways

for the biosynthesis of all biomass building blocks. GEMs are capable of synthesizing biomass, which means that the synthesis pathways for all biomass building blocks (BBBs) can carry flux. However, due to the degrees of freedom that the system has, it is almost impossible to determine all the enzymes contributing for the synthesis of each biomass building block, and the precursor metabolites that these BBBs are synthesized from by simply solving the system for a linear cellular objective. Also the textbook pathways cannot explain the complexity in the network, since they do not account for the nested parts of the pathways. In literature, there are reported values for the stoichiometric expenditure of precursor metabolites for some BBBs, for instance in classical textbooks, such as Neidhardt[21], but these estimations are not complete as they do not account for the latest knowledge in GEMs and for multiple alternative routes that can be used to synthesize the same BBBs. To overcome these limitations, we have developed lumpGEM, a method, which enumerates possible subnetworks for the synthesis of every biomass building block from core carbon metabolites. In this way, we have re-defined the possible pathways for BBBs and revealed the *nested* parts of these pathways that cannot be deduced systematically through growth maximization simulations or from databases such as KEGG[22], EcoCyc[23], and SEED[24]. Apart from enumerating synthesis pathways for BBBs, and re-defining the biosynthetic pathways, lumpGEM also reduces the complexity of GEMs by lumping the generated subnetworks for each BBB, and representing them through 1 overall reaction, which is stoichiometrically equivalent to the calculated subnetwork for the generated flux distribution. This gives us the opportunity to focus on certain parts metabolism and reduce the complexity emerges from different parts of the network by lumping.

Bottom-up approach allows us to build high quality models, such as GEMs, based on detailed mechanistic knowledge with well-established pipelines[25]. Top- down approach, on the other hand, extracts information from accumulated data and knowledge in the literature to build models. Top-down models are usually hypothesis-driven[26] and focus on certain functions, rather than having a global approach, and unlike bottom-up models, have small sizes. Keng et al. [27] discussed the need to build models having the bottom-up quality within the size of top-down models. In Chapter 5, we have devel-

oped redGEM, a tool that builds reduced core models around the part of the metabolism under interest. redGEM uses the bottom-up GEM models as input and generates reduced models while keeping the mechanistic quality of GEMs. Moreover, redGEM uses lumpGEM to generate lumped reactions for all biomass building blocks from the core metabolites. Final output of redGEM is a reduced core model that is consistent with its GEM in terms of biomass yields, by-product formation, allowable flux ranges and essential genes and reactions. These core models are valuable platforms for various applications, from C13 labeling experiments[28] to TFA to kinetic models since they offer manageable model sizes for such studies. Moreover, reduced models can be used as systematic comparison tools to understand the fundamental differences of metabolic and biosynthetic capabilities between organisms. In Chapter 6, we focused on the central carbon networks generated by redGEM and lumpGEM for 3 different organisms, *P. putida, S. cerevisiae* and human, and highlighted the differences between them. Moreover, we generated subnetworks and their lumped reactions for the 3 organisms and showed their similarities and differences.

Integration of thermodynamics and reduction of GEMs are two very powerful methods to reduce the solution space and the complexity of the metabolic networks. However they do not characterize the still big solution space. The source of this complexity of solution spaces can be classified into two main categories, *i) the flexibility in the absolute amount of the flux through a reaction* and *ii) the directionality of the reaction.* The latter emerges from bidirectional reactions (BDRs) in metabolic networks. To enumerate and slice the solution spaces into subspaces through BDRs, in Chapter 7, we have developed a method called Flux Directionality Profile Analysis (FDPA), and enumerated all possible flux directionality profiles (FDP) for *E. coli* grown under aerobic conditions with glucose and succinate as sole carbon sources. Each of these FDPs represents a certain physiology, and has different characteristics in terms of theoretical maximum growth, by-product formation etc. Two of the most interesting studies that FDPA allows are the study of flexibility that emerges from absolute amount of fluxes through sampling the sub-solution space unique to an FDP and displacement of reactions from thermodynamic equilibrium.

The metabolism of certain organisms, such as *E. coli* or *S. cerevisiae* has been well characterized and they have been studied for decades. However, even for these organisms, there are metabolic gaps and unknown biochemical reactions that are characterized in every year. For instance, databases that are updated annually, such as KEGG, are adding new metabolic enzymes and reactions in every update. These newly added reactions can represent a new biochemistry or can be new reactions catalysed by a known biochemistry rule. Methods to predict these reactions and many more possible candidate reactions have been proposed in the last 15 years. BNICE.ch (Biochemical Network Integrated Computational Explorer)[29] is a powerful tool that can reproduce all known reactions with a defined biochemistry in KEGG database and is capable of proposing novel biochemical reactions with manually built reaction rules through known biochemistry. In Chapter 8, we used BNICE to explore the metabolic capabilities of *E. coli* by generating all known and novel reactions around the central carbon metabolism of *E. coli* defined as Glycolysis, TCA Cycle, Pentose Phosphate Pathway, Pyruvate Metabolism and Glyoxylate shunt. BNICE.ch also brings a new kind of multiplicity and complexity for metabolic modelling purposes, since it explores the whole space of biochemistry and proposes thousands of novel reactions. In Chapter 8, we also aimed in reducing this complexity. By merging the native metabolic network of *E. coli* and the reactions proposed by BNICE.ch, we built a *super* network, and explored its metabolic capabilities. We revealed sets of novel or known non-native *E. coli* reactions in this *super* network that can increase the biomass yield of the organism and exhaustively generated all possible thermodynamically feasible sets. We also performed a gap filling analysis for the dead-end metabolites in the native network, and identified possible biotransformations for these compounds throughout the *super* network.

In this thesis, we targeted navigating and managing the complexity of metabolic networks for a better understanding of the cellular physiology, and to guide biotechnological applications. We built tools that can reduce this complexity, and proposed novel methods to analyse both reduced and genome scale networks. These applications are not developed for specific organisms, and can be applied to all GEMs and top-down built metabolic networks, as we demonstrated on *P. putida,* yeast, and human. Also, by

BNICE.ch, we explored the metabolic potentials of *E. coli* by using known biochemistry rules, which can also be performed on any organism. We believe that these tools/methods will help us to develop a better, more comprehensive view on one of the most important component of life, metabolism.

Apart from the work discussed in this thesis, I also contributed for the development and refinement of different tools that is used in our group, LCSB. I have developed a pipeline that evaluates pathways for the production of target chemicals. This pipeline generates thermodynamic information for all the metabolites and reactions on these pathways using GCM, merges them with GEMs, performs FBA/TFA analysis and ranks them based on the yield for the target chemicals. The method can evaluate more than ~15000 pathways per day. Moreover, I merged the traditional gap-filling analysis with TFA and built a method that can perform thermodynamically feasible gap filling analysis per biomass building block, or any metabolite that is targeted. I also contributed to the refinement of in-house TFA toolkit AGADOR and built new thermodynamics-based methods for different analysis, such as gene essentiality and minimal media analysis, which are already used in a publication [30].

Chapter 2 - Heading in the right direction: Thermodynamics-based network analysis and pathway engineering

2.1 Introduction

In the last 15 years, the number of annotated genome sequences has grown tremendously, and this has led to reconstruction of genome scale metabolic models (GEM) for many organisms, from unicellular prokaryotes to higher organisms such as mouse and human [31]. These metabolic models are *in silico* representations of all biochemical reactions that take place in the cell. Through various methods, such as Flux Balance Analysis (FBA), different phenotypes of organisms can be simulated and analysed [32,33]. Directionalities and allowable flux ranges for metabolic reactions are the main constraints that delineate the boundaries for GEMs. The two most important uses of thermodynamics-based analysis of metabolic networks are the determination of reaction directionality and the estimation of how far from, or close to, equilibrium the reactions in the network operate.

In most of the metabolic flux balance studies that discuss and analyse reaction thermodynamics and energetics, the authors consider the reactions as irreversible (unidirectional) based on the standard Gibbs free energy of reaction. Soh and Hatzimanikatis [13,34] suggested differentiating between "reaction directionality" and the commonly used term "reaction reversibility". Reaction reversibility is a *kinetic* property of the enzyme and it is used to denote that the enzymes are able to catalyse the reactions in both directions, i.e. the *forward* and *backward* reactions. If an enzyme is catalytically reversi-

ble, then the directionality of the reaction depends on the displacement of the reaction from thermodynamic equilibrium. In the context of a metabolic network, with or without thermodynamic constraints, a reversible reaction can be either *bidirectional*, i.e. it is able to operate in both the forward and reverse directions, or *unidirectional*, i.e., it can operate only in one of the directions. The catalytic reversibility is an enzyme property that depends on the enzyme amino acid sequence, and therefore it can be different between organisms. The information about catalytic reversibility is available for a relatively small number of the enzymes in the biological databases and for a very small number of organisms. Therefore, by determining the reaction directionality, thermodynamic constraints provide important information that substitute for the lack of information about reaction reversibility.

Three main approaches have been used for the introduction of thermodynamic constraints (network thermodynamics): (i) the energy balance analysis (EBA)[35], (ii) the network-embedded thermodynamic analysis (NET analysis) [36,37], and (iii) the thermodynamics-based flux analysis (TFA), which has been also called metabolic flux analysis (TMFA)[14] or thermodynamics-based flux balance analysis (TFBA)[13]. All three methods introduce a new set of constraints that enforce the reactions fluxes to operate within the feasible bounds of energy constraints. The general EBA problems constrain the directionality and bounds of the fluxes using the value of the Gibbs free energy, either as a constant or as continuous variable within defined ranges. NET analysis and TFA constrain also the fluxes using the value of the Gibbs free energy but as a linear function of the logarithms of the metabolite concentrations (or activities). However, NET analysis requires a predetermination of the directionality of the fluxes, and the thermodynamic constraints determine if the flux is feasible in the defined direction and what are the thermodynamically feasible concentration ranges. The TFA considers initially all catalytically reversible fluxes as thermodynamically bidirectional and employs a mixed-integer linear programming formulation that accounts for concentration ranges and it computes the flux directionality based on the thermodynamically feasible concentration profiles. Therefore TFA introduces the minimum bias about reaction directionality and it simultaneously computes thermodynamically feasible flux *and* concentration

ranges. Moreover, the EBA and NET analysis formulations represent special cases of the TFA formulation. Hence, we believe that any analysis that uses thermodynamic constraints should apply TFA, or a similar formulation in order to avoid incomplete or false predictions about the properties of the network.

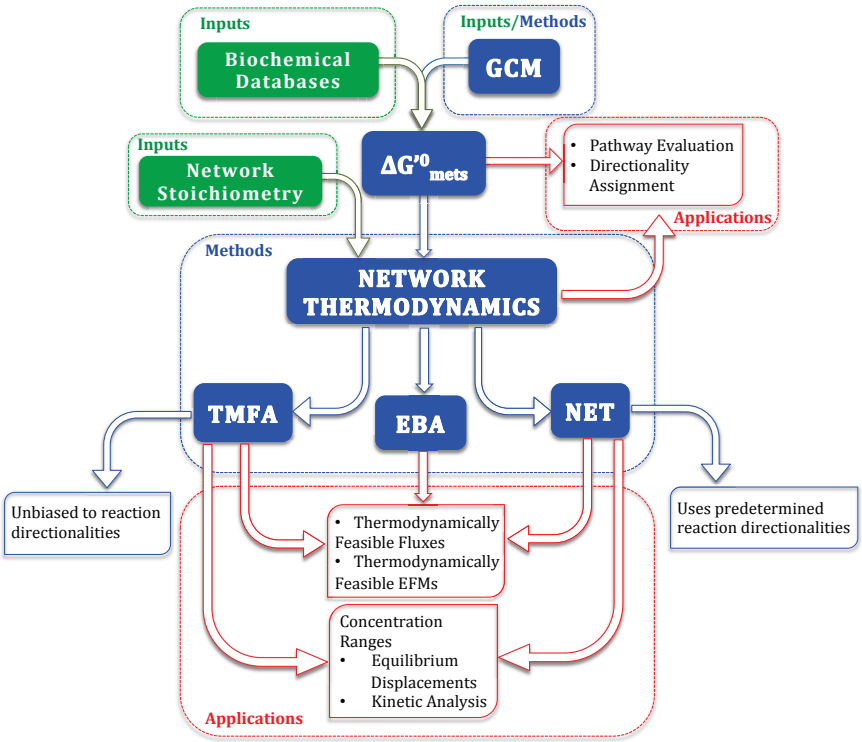

Figure 2-1**:** Applications of thermodynamics in metabolic networks. Standard Gibbs free energy values can be used to evaluate pathways, and to assign directionalities in GEMs. Integration of network-thermodynamics allows computation of thermodynamically feasible flux profiles and EFMs and feasible concentration ranges. These ranges can be further used for estimating the equilibrium displacements and building kinetic models of metabolic networks.

We review here the recent publications that have applied thermodynamic constraints in metabolic flux balance analysis. The major applications of thermodynamics have benefited the study of metabolism through three main uses: (i) the application of thermodynamic constraints to assign directionalities and thus constraint the allowable flux space and improve the predictions of metabolic modelling; (ii) the evaluation of the feasibility

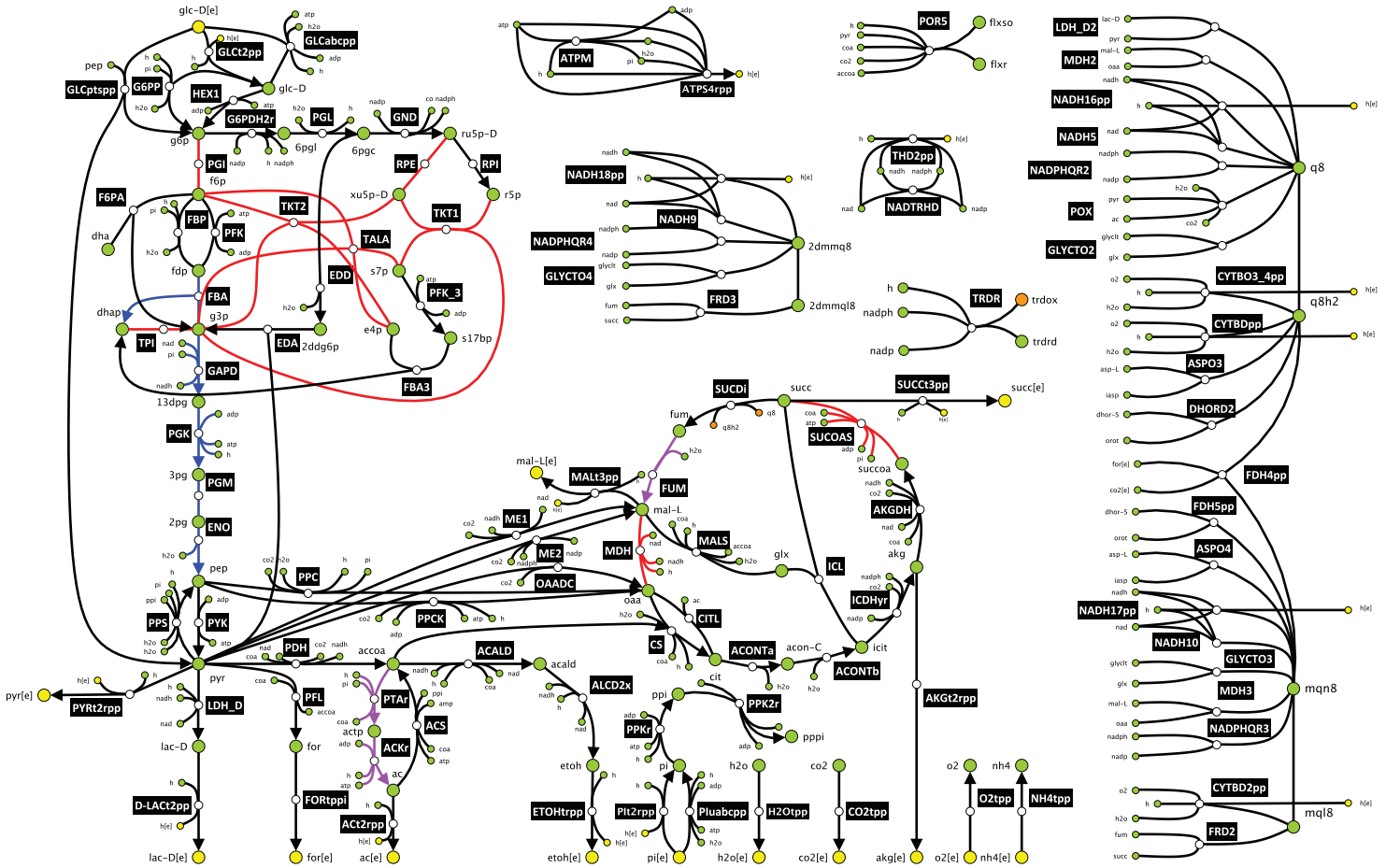

Figure 2-2: Bidirectional reactions in reduced *E. coli* network. Without any thermodynamic constraints, the reactions colored with blue, red and purple are assigned as bidirectional by FBA. By applying TFA on the network with wide metabolite concentration ranges, the reactions colored with blue and red are bidirectional. Integration of metabolomics data [38] to TFA constrains 3 more reaction as unidirectional (colored with red). Using only standard Gibbs reaction values for directionality assignment cannot capture the impact of concentration ranges on directionality of reactions.

of synthetic and metabolically engineered pathways; and (iii) the integration of metabolomics data into metabolic models and their analysis and interpretation in the context of metabolic networks (Figure 2.2).

2.2 Assigning directionality based on Gibbs free energy of reactions in GEM

Directionalities of metabolic fluxes in GEMs have significant impact on network properties, such as yield on biomass, gene/reaction essentiality. Directionality specifications in GEMs are based on literature, databases, biochemical textbooks and information from similar organisms. If there is no available information about the catalytic reversibility of an enzyme, the corresponding metabolic reaction is defined as *bidirectional* in the network. The sole systematic approach then to account for the directionality for these reactions is integration of thermodynamics into metabolic networks. In a recent study, Dreyfuss et. al [39] reconstructed the genome scale metabolic model of *Neurospora crassa*, and by using the Gibbs free energy of reactions through the Group Contribution Method (GCM) [40-42], they constrained 1046 of 1374 metabolic reactions as unidirectional. However, they did not take into account the contribution of the activities to the Gibbs free energy of reactions, and they determined the directionality just by using the standard values. Pitkanen et. al [43] followed the same path, and they assigned directionalities to the reactions in 49 fungal species' GEMs by using eQuilibrator[44]. Assigning directionality based on the Gibbs free energy approach is becoming more and more popular, and has been applied to many metabolic networks reconstructions [45-51]. All these calculations are done without a systematic integration of thermodynamics constraints. A systematic approach provides a more accurate estimation on the directionality of reactions in metabolic networks rather than using only standard the Gibbs free energy of reactions (Figure 2).

Thermodynamic constraints can also be used to test the consistency of pre-determined directionality in GEMs. By utilizing the NET analysis method, Martinez et. al [52], identified 319 unidirectional reactions in Recon 1, human metabolic network model. 306 of these reactions were already set as unidirectional in GEM, whereas 13 of them emerged

as new unidirectionality constraints. They also revealed thermodynamically infeasible internal loops, and removed these biologically meaningless futile cycles. Moreover, they concluded that organisms use alternative methods in order to overcome the thermodynamic constraints, such as substrate channelling or coupling with ATP hydrolysis. In this study as in all the current studies that use GCM data, the standard Gibbs free energy values used were estimated for 25 °C, while the human body temperature is typically 12 degrees higher. This points to one of the important needs for expanding the GCM methods for broader ranges of temperature and pressure that can be used for the analysis of systems like human and extremophiles.

2.3 Improving model predictions by integrating thermodynamics with metabolic models

Accurate prediction of the observed phenotypes is one of the main evaluation criteria of the quality of metabolic models. Thermodynamics, along with other constraints, such as mRNA and protein expression data, are widely used to improve the prediction capacity of metabolic networks. For instance, by applying thermodynamics constraint in a systematic manner by TFA, Soh et. al [27] showed that the optimum specific growth rate of *Saccharomyces cerevisiae* drops from 1/hr to 0.42/hr with 15.33mmol/gDWhr glucose uptake rate, and the experimentally observed value was 0.35/hr. In another study, Schulz and Qutub proposed a method (corsoFBA) to integrate a 'thermodynamics cost' to metabolic networks, similar to EBA analysis, while optimizing the protein cost to study sub-optimal growth phenotypes. They succeeded in capturing the metabolic fluxes at sub-optimal growth from experimental data of *E. coli* at different dilution rates [53].

NET analysis has been also used to test the thermodynamic feasibility of given directionalities in metabolic networks. De Martino and colleagues [54] proposed a method that attempts to overcome the NET analysis' requirement of pre-assigned reaction directionalities. They used their method to estimate the metabolite concentration ranges for human red blood cells and to identify thermodynamically infeasible loops in *E. coli*.

However, their formulation cannot guarantee a globally optimal solution and cannot enumerate alternative optimal solutions as it is done by the TFA.

Another study combined NET, and experimental, analysis of *E. coli* [55]. The authors revealed that the Acetyl-CoA C-acetyltransferase is thermodynamically infeasible in acetoacetylCoA synthesis direction. Under anaerobic and aerobic conditions, this infeasibility results in a 2.9% and 1.1% reduction in growth, respectively.

Orman et. al [56] used EBA analysis to study the behaviour of the perfused livers under fed and fasted states. They applied EBA on certain parts of the network, such as pathways, and assumed that the overall dissipation of these pathways must be equal or greater than 0. The result was a significantly reduced solution space.

These studies further demonstrate the benefit of thermodynamic constraints and it will be interesting to compare their results with the TFA analysis, which is not subject to preassigned reaction directionalities.

2.4 Thermodynamics methods in Systems and Synthetic Biology tools

Integrating thermodynamic information, either through Gibbs free energy of formation, or in a network manner such as TFA, has become one of the most important features of systems biology and metabolic engineering for the analysis and design of synthetic metabolic pathways.

Pathway design tools [57-61] use the standard Gibbs free energy of reactions estimated through GCM[40] to prune the set of *de novo* generated pathways and to retain only the thermodynamically feasible ones. However, they do not include any systematic network-thermodynamics approach to account for the effect of concentrations of metabolites on the overall thermodynamic feasibility of the pathways. The necessity of adjusting the estimated standard Gibbs free energies to physiological conditions (metabolite concentrations, pH, and ionic strength) is discussed in [62,63]. These studies also demonstrate that the number of feasible pathways can be reduced significantly if thermodynamic constraints are applied. The pruning that is based only on the standard

Gibbs free energy, and using methods like EBA, is very conservative. Many pathways that would be discarded as infeasible using standard Gibbs free energy could become feasible using a framework like TFA, which allows for adjustment of the free energy to physiological conditions.

The use of thermodynamics in systems biology tools is not limited to pathway evaluations and can also be used to provide ranges for the flux values. Thermodynamic Optimum Searching (TOS) [64] aims to calculate the thermodynamically optimal flux solution by minimizing the magnitude of Gibbs free energy change and maximizing the entropy production with an EBA type analysis. Muller and Buckmayr [65] propose a similar method with improved computation time. However, these methods are based on the formulation of EBA, and therefore their results are biased in the pre-selection of Gibbs free energy bounds. It remains to be shown how the reformulation of these methods can integrate metabolite concentrations as variables and how they can reduce the computational cost associated with such integration.

2.5 Thermodynamically Feasible Elementary Flux Modes

Elementary Flux Modes (EFMs) analysis, which characterizes the allowable steady state fluxes for a metabolic network [66], has been extensively used to investigate the capabilities of metabolic networks. However, even a small network can have millions of EFMs, and this necessitates the usage of methods to characterize and eliminate biologically irrelevant EFMs. Gerstl et. al [67] developed a framework to identify thermodynamically feasible EFMs by utilizing NET approach. This method reduced the number of EFMs significantly by eliminating the thermodynamically infeasible EFMs. A very similar approach was followed by Jol et. al [68], in which they calculated 71 million EFMs of *S. cerevisiae* metabolic network, and through the NET analysis, they concluded that 56% of the EFMs are thermodynamically feasible. A method that integrates thermodynamics into EFM analysis with improved efficiency has been also developed[69]. These are very promising results and new model formulations and algorithms can lead to significant reductions in the number of EFMs, removing one of the main limitations for their broader applicability.

2.6 Thermodynamically feasible concentration ranges, kinetic modeling and metabolomics

We can further use network thermodynamics to integrate metabolomics data in metabolic network models and to evaluate the consistency of these data with the metabolic flux profiles. Since metabolite concentrations determine the Gibbs free energy of a reaction, the concentrations of the reactants must be consistent with the flux directionality. Network thermodynamics can also be used to validate the experimental results of *in silico* predictions, such as measured metabolite concentrations [70], or to explain the variations in concentration levels under different growth conditions [71].

Soh and Hatzimanikatis [13] defined the *thermodynamic space of the network* as the space of *the thermodynamically allowable metabolite concentrations* and the space of *the reaction displacements from equilibrium*, which are constrained by the Gibbs free of the reactions in the network. The thermodynamic space of a metabolic network can be characterized and analysed through the sampling of the concentrations of the metabolites and/or the corresponding Gibbs energies of reactions using a TFA or a NET analysis formulation. Using such approaches Soh et al [27] have derived for the first time the displacement from equilibrium for a metabolic network in yeast, which includes the central carbon pathways and the cytosolic and mitochondrial electron transport chains. A similar approach was also used by Birkenmeier et al [72] to generate thermodynamically feasible pathways, by sampling the metabolite concentrations by NET analysis approach. They analysed the glycerol biosynthetic pathway of *S. cerevisiae* without a *priori* knowledge of specific enzyme kinetic rate laws and parameters. They concluded that the pathway is primarily controlled by glycerol-3-phosphate dehydrogenase enzyme that operates far from equilibrium; which was previously proven experimentally.

Thermodynamically feasible steady state concentration profiles can be further used for kinetic analysis of organisms. Chakrabarti et al [73] developed a method to build kinetic models for genome scale reconstructions that takes into account all the stoichiometric and thermodynamic constraints of the flux balance models. In this work, they calculated the thermodynamic space in a metabolic network of the 146 reactions and 90 metabo-

lites that describe the central carbon metabolism and electron transport in *E. coli* using the TFA formulation. For each sampled set of the thermodynamic allowable concentrations, they next calculated thousands of kinetic models that were stable and consistent with the allowable concentrations and flux profiles. The efficiency of the method in building kinetic models of such size and quality depends strongly on the proper choice of thermodynamically feasible concentrations early in the model-building process. Similar studies by Milo and colleagues [74,75] have shown that the integration of thermodynamic constraints and the decomposition of the rate expressions between the kinetic and the thermodynamic terms can improve the process of building kinetic models and provide important insights into the analysis of complex kinetic models.

2.7 Conclusions

While thermodynamics have been used in many studies, their use is still limited relative to the enormous field of metabolic modelling. Plants, due to their importance for energy capture, and extremophiles, due to their non-standard bioenergetics properties, are very promising organisms for biotechnology but the study of their bioenergetics is very challenging. Simons et al [76] reconstructed a metabolic model of a maize leaf and they used the Gibbs free energy of reactions to remove thermodynamically infeasible cycles. PlantSEED [77], a comprehensive computational environment that focuses on plant metabolism, includes in its database thermodynamic properties of the metabolites and reactions, and it defines cellular subsystems and compartments based on metabolic reconstructions for plants. Introduction of thermodynamic constraints in metabolic models that include multiple compartments is a challenging task that will require careful formulation of the thermodynamic constraints for the transport reactions [78]. TFA formulations for such networks can reveal important properties of the energy metabolism and the bioenergetics properties of plants and other multi-compartmental organisms.

A recent study [79] on the adaptation of *Saccharomyces* species to different temperatures, accounted for the Gibbs free energies of reactions under different conditions, and predicted the metabolic changes that keep cells alive, such as increased glycerol accu-

mulation. To perform detailed study on organisms under such harsh conditions, it is essential to have accurate predictions for thermodynamics properties at high and low temperatures or high pressures. Recently, with the integration of quantum chemistry, this issue has been addressed, and new methods are proposed to achieve this goal [80,81].

Regardless of the organism under study, all future studies will require development of methods for: (i) the identification and reduction of number of bidirectional reactions, and (ii) the identification of the metabolites which, if measured, would allow us to estimate the displacement of reactions from thermodynamic equilibrium with higher confidence. Using methods like EBA and TFA we can identify the number of the bidirectional reactions. However, alternative combinations of flux directionalities can grow enormously as the number bidirectional reactions increases. Methods that can rank the alternative flux directionalities and the associated flux profiles can provide a systematic analysis of cellular physiology. Similarly the identification of the most informative metabolites will be very important for studies in metabolomics, physiology and bioenergetics. These methods should be able to handle large- to genome-scale networks and to account for the uncertainty in the input data. We expect that developments in these areas will further expand the scope and the usefulness of network thermodynamics.

This Chapter is published with the following details:

Ataman M, Hatzimanikatis V: ***Heading in the right direction: thermodynamics-based network analysis and pathway engineering****. Current opinion in biotechnology 2015,* ***36****:176-182.*

Chapter 3 – *In silico* Studies on Bioenergetics of Aerobically grown *E. coli*

3.1 Introduction

Organisms utilize the energy stored in the carbon sources such as sugars, lipids, etc. to maintain a lifecycle [82]. For many organisms, the source of this energy is obtained from oxidation of organic compounds. The interconversion of this internal energy is a one of the key points to understand the behaviour of organisms. The energy obtained from this oxidation is captured in other organic molecules, such as ATP, NADH or reduced flavin [83]. These molecules are then oxidized or broken down in order to catalyse the biochemical transformations that are necessary for the survival of the organism [84]. Comparison between the energy stored in biochemical molecules and the energy that organisms obtain from those molecules shows that the efficiency of energy capturing is not 100 per cent [85], and significant amount of energy is lost during this process. This energy is dissipated to the environment as internal energy captured in by-products and as heat. The latter helps the cell to maintain a certain internal temperature, and to avoid overall chemical equilibrium [86].

In cells, the main metabolite that captures the internal energy of the carbon source is ATP. The production of ATP, under aerobic conditions, is directly related with the oxidation/reduction reactions of a subsystem called Electron Transport Chains (ETC). ETC is the pathway that is composed of a series of oxidation and reduction processes, which terminates with the reduction of molecular oxygen to water [87] under oxygenated environments for many organisms. This pathway is coupled to the synthesis of ATP (oxidative phosphorylation) via the enzymes translocating protons to extracellular or

periplasmic medium, thus creating a proton motive force (*pmf*) [88]. P:O Ratio, which represents the ATP production per oxygen atom reduced in Electron Transport Chains (ETC), has been defined as the stoichiometric efficiency for organisms in terms of ATP synthesis [82,83].

Over decades, researchers studied in details on the thermodynamic efficiency of oxidation of carbon sources, and the synthesis of ATP via oxidative phosphorylation, separately. With non-equilibrium thermodynamics, the relation between the growth rate and total Gibbs free energy dissipation of the organisms has been shown extensively [84,89,90]. Moreover, in these studies, mathematical models to predict the experimentally observed growth rates have been built. In such studies, the yield for ATP production is also included, however these models are very simple compared to the complexity that exists in cellular networks and ATP yield is included in the models usually as constants. The stoichiometric ATP synthesis capacity of the organisms is studied mainly through P:O Ratio. Experimentally, stoichiometric P:O Ratios have been determined for different organisms [91-94], under different conditions, such as alternative carbon sources, or nitrogen limitations. Also, these stoichiometric values have been utilized in a few *in silico* analysis [18,95,96]. However, up to date, to our knowledge, P:O Ratio, which accounts for the efficiency of organisms to capture internal energy obtained from oxidizing its carbon source, and Gibbs free energy dissipation, the energy that is lost during this process have not been studied jointly in metabolic networks.

In this chapter, we focused on the bioenergetics of cellular networks, by utilizing genome scale metabolic reconstruction of *E. coli* [6] constrained with network thermodynamics TFA [14] to study its P:O Ratio and Gibbs free energy dissipation behaviour, and to reveal how those two bioenergetics concepts are interrelated. Moreover, we focused on the behaviour of by-product formation of *E. coli* under aerobic conditions, and its relation with the thermodynamic efficiency of this organism. We concluded that *E. coli* interconverts the internal energy obtained from its sources, in this specific case glucose, to biomass with highest efficiency at the experimentally observed optimum P:O Ratio, and this point coincides with highest theoretical specific growth rate. Also, we identified the enzymatic biotransformations, which are the main components of the dissipated

free energy, and we concluded that these enzymes are mainly elements of ETC pathways. Finally, we showed that acetate, as the main by-product for *E. coli*, is also the most efficient by-product formation route, to minimize the Gibbs free energy dissipation to the environment.

3.2 Materials and Methods

For simulations, we used the latest genome scale model of *E. coli* iOJ1366, which is composed of 2253 enzymatic reactions and 1138 unique metabolites, along 2 compartments, periplasm and cytosol. We thermodynamically curated the model with an in-built toolbox, called AGADOR, which uses the TFA[13,14] formulation. In this work, we focused on the aerobically grown *E. coli* with glucose as the sole carbon source. Therefore we used the reported values for uptakes and secretion under these conditions from literature [97]. TFA formulation allows the usage of metabolite concentrations, both intracellular and extracellular; hence we integrated *ec (energy charge), crc (catabolic reduction charge) and arc (anabolic reduction charge)* to the model by integrating the metabolomics data from another study from Bennett et al. [38]. We directly constrained the absolute metabolite concentrations of ATP, ADP, AMP, NAD^+, NADH, $NADP^+$, and NADPH. In order to account for the extracellular metabolite concentrations such as sulphate, phosphate, ammonia, carbon dioxide and oxygen, we integrated data from the study from Henry et al. [14].

We simulated each P:O point by setting the ratio between the ATP synthase flux and the reactions that use oxygen in Electron Transport Chains of aerobically grown *E. coli*, three reactions catalysed by cytochrome enzymes (bd-1, bd-2 and bo3). We followed the steps listed below to simulate different P:O ratio values for Flux Balance Analysis:

1. We create a pseudo metabolite.
2. We create a new constraint for the balancing of this metabolites as follows:
    a. We add the pseudo metabolite into ATP synthase reaction with the stoichiometric coefficient 1 (as product) and cytochrome reactions with the coefficient γ (as reactant).
    b. We create the mass balance constraint for this metabolite:

i. Flux_ATP_synthase – γ (Flux_3_Cytochromes) = 0.

c. We generate different solutions by varying γ incrementally.

We used the same approach to simulate P:O ratio in TFA formulation, and perform simulations at different P:O ratios to study its effect on the metabolic network behaviour under mass and thermodynamic constraints. The only difference is that in TFA formulation, we can add the constraint for P:O ratio without introducing a pseudo metabolite in the formulation.

To calculate the Gibbs free energy dissipation of the cell (ϕ), we utilized the Gibbs free energy of reactions, and the corresponding flux values. The formula that we used to calculate the Gibbs free energy dissipation is as follows:

$$\phi = \sum_j \left(v_j \Delta G'^0_{rxn,j} + v_j RT \sum_i n ln(c_i) \right) \tag{3.1}$$

In this equation, v_j represents the flux through jth reaction, c_i is the concentration of metabolite *i* in the reaction v_j, with the stoichiometric coefficient as *'n'*. If we write this equation by inserting the stoichiometric matrix, we reach to the following representation:

$$\phi = v^T \Delta G'^0 + RT(v^T N^T) ln(c) \tag{3.2}$$

In this representation, v^T is the transpose flux profile vector, N^T stands for the transpose of stoichiometric matrix, *'c'* is the vector that keeps the concentration of each metabolite in the system and $\Delta G'^0$ is the vector that keep the transformed standard Gibbs free energy of reaction for each reaction in the network. Pseudo steady state assumption assumes no accumulation of any metabolites in cells; hence the mass balance for all metabolites in Flux Balance analysis is represented through the following constraint:

$$N.v = 0 \tag{3.3}$$

If we split the stoichiometric matrix into 2 parts, namely enzymatic reactions (N^{enz}), and boundary reactions (N^{bnd}), $v^{T,bnd}N^{T,bnd}$ will not have a contribution to ϕ since boundary reactions are not real reactions, but just define the systems boundaries, therefore they do not have a $\Delta G'^{0}_{rxn}$. If we write the Gibbs free energy dissipation term with Formula 3.2, the term $(v^T N^{enz})ln(c)$ will have 0 contribution for cytosolic and periplasmic metabolites, but will have contribution for the extracellular metabolites, since these metabolites will not be balanced by boundary reactions. Therefore media composition will have a direct effect on the Gibbs free energy dissipation of the organism, along with its indirect effect by changing the possible flux distributions in the metabolic network.

The values reported in this paper for Gibbs free energy dissipation is adjusted for production of 1 Carbon mole of biomass. In order to calculate the Gibbs free energy dissipation for this term, we used the following representation:

$$\phi(1\ Carbon\ mole\ of\ biomass) = \frac{\phi(total)}{Carbon\ mole\ in\ biomass * Specific\ Growth\ rate} \tag{3.4}$$

$$Carbon\ mole\ in\ biomass = \sum_{i}^{\#\ of\ BBB} n_i * BBB_{i,\#\ of\ Carbon} \tag{3.5}$$

in where n_i is the stoichiometric coefficient of BBB_i in the biomass formulation.

3.3 Results and Discussion

The studies about bioenergetics of metabolic networks revolve around respiratory chains, mainly oxidoreductases, dehydrogenases, ATP synthase enzymes and transport reactions across membranes. All these reactions are parts of GEMs, with well-defined stoichiometry. Moreover, TFA allows us to focus on bioenergetics of the whole cell with a systems perspective through detailed thermodynamics descriptions of enzymatic reactions and biochemical transports processes[78].

Firstly, we focused on the stoichiometric efficiency of ATP synthesis of *E. coli* under aerobic conditions with glucose as the sole carbon source. Experimentally, P:O ratio is defined as the number of ATP molecule produced per oxygen atom uptaken by the cell [92]. Slightly different, we define P:O ratio as the flux through ATP synthase reaction and the total flux through reactions catalysed by Cytochrome enzyme family (Table 3.1), which utilizes oxygen to create proton motive force (pmf) across the compartments. Since some other enzymes in the organism can use molecular oxygen, this representation captures the coupling between the oxygen usage as terminal electron acceptor and ATP synthesis. Moreover, the protons that are pumped from cytoplasm to periplasm via cytochromes are coupled to ATP synthase reaction; the amount of reduced oxygen is directly related to the amount of ATP that the cell synthesizes via oxidative phosphorylation.

$$\mathrm{P{:}O} = \frac{v_{\text{ATP snythase}}}{2.\sum_i v_{\text{cyt,i}}} \tag{3.6}$$

The denominator is multiplied by 2, since the oxygen in P:O definition is based on atomic oxygen.

Table 3-1 Stoichiometry of the each Cytochrome Oxidoreductases and ATP synthase reactions in the *E. coli* metabolic network.

Enzymes	Enzymatic Reaction Formulas
CYT_{bd1}	2 H^+(cytosolic) +Ubiquinol8+0.5O_2 → H2O + Ubiquinone8 + 2 H^+(periplasmic)
CYT_{bd2}	2 H^+(cytosolic) + Menaquinol8 + 0.5O_2 → H2O + Menaquinone8 + 2 H^+(periplasmic)
CYT_{bo3}	4 H^+(cytosolic) + Ubiquinol8 + 0.5O_2 → H2O + Ubiquinone8 + 4 H^+(periplasmic)
ATP synth	ADP + Phosphate + 4 H^+(periplasmic) → ATP + 3 H^+(cytosolic) + H2O

The reported values for P:O ratio of *E. coli* in the literature are close to each other, 1.5 ± 0.1 [96] and 1.49 ± 0.26 [91] under aerobic, glucose minimal media, which they define a range, rather than a single value. Moreover, Bekker et al. showed that P:O ratios can vary due to the alternative enzymes in the aerobic respiratory chains of *E. coli*[98]. Difference in the stoichiometry of these enzymes, mainly the number of pumped out protons contributes to the flexibility and variability in observed P:O ratios.

Metabolic models allow creating constraints that can represent splitting points between any parts of the networks. P:O ratio is not a splitting point, however different P:O ratios can be represented as splitting points between ATP synthase reaction and the total flux through cytochrome reactions in Formula 3.6 (Materials and Methods).

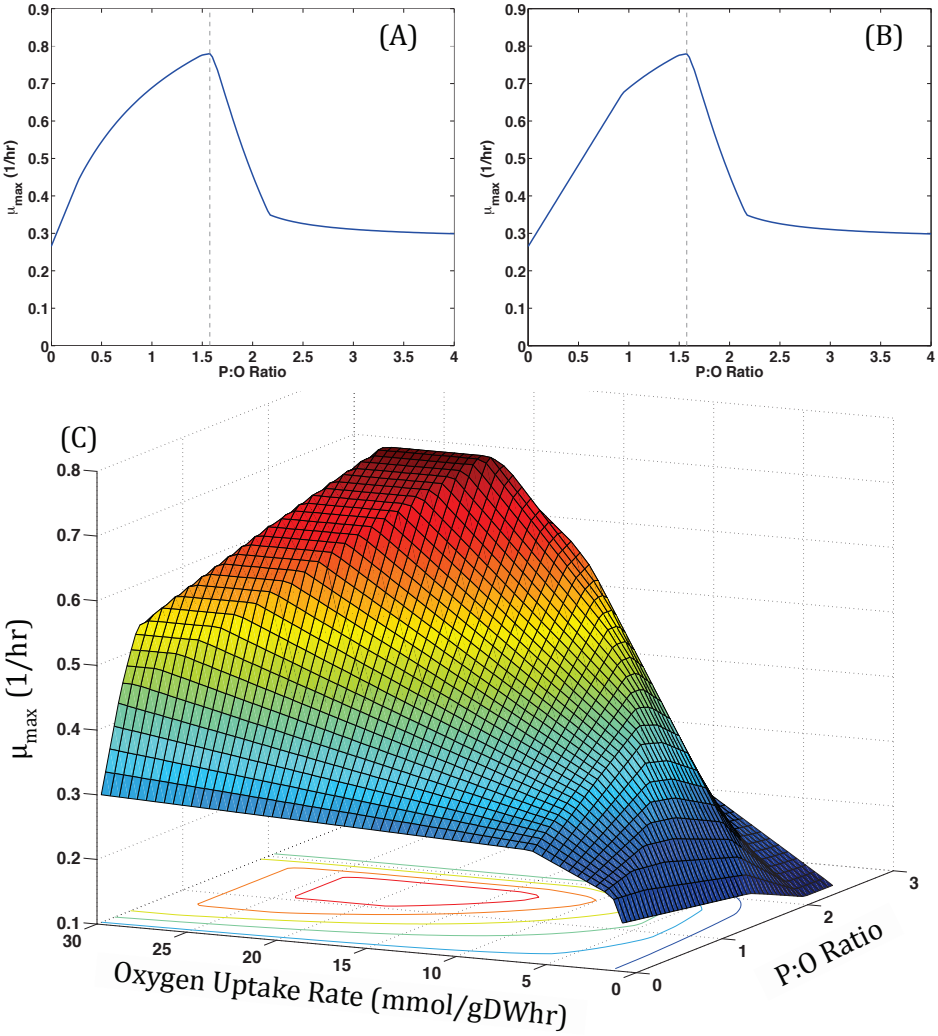

Figure 3-1:Feasible P:O range and corresponding specific growth rates. A) Flux Balance Analysis (FBA) results. C) Same analysis repeated with thermodynamic constraint implied by TFA. C) Feasible P:O range under different Oxygen Uptake Rates with TFA. By changing the Oxygen Uptake Rate of *E. coli* metabolism from 1 mmol/gDWhr to 30 mmol/gDWhr, feasible P:O ranges for each uptake rate is determined. P:O value is iteratively increased by a step of 0.001 from 0 to 2 for each oxygen uptake rate, and correspond-

ing specific growth rates are determined. The plateau shows the adaptation of the organism to different oxygenation states, and its ability to survive under different environmental conditions.

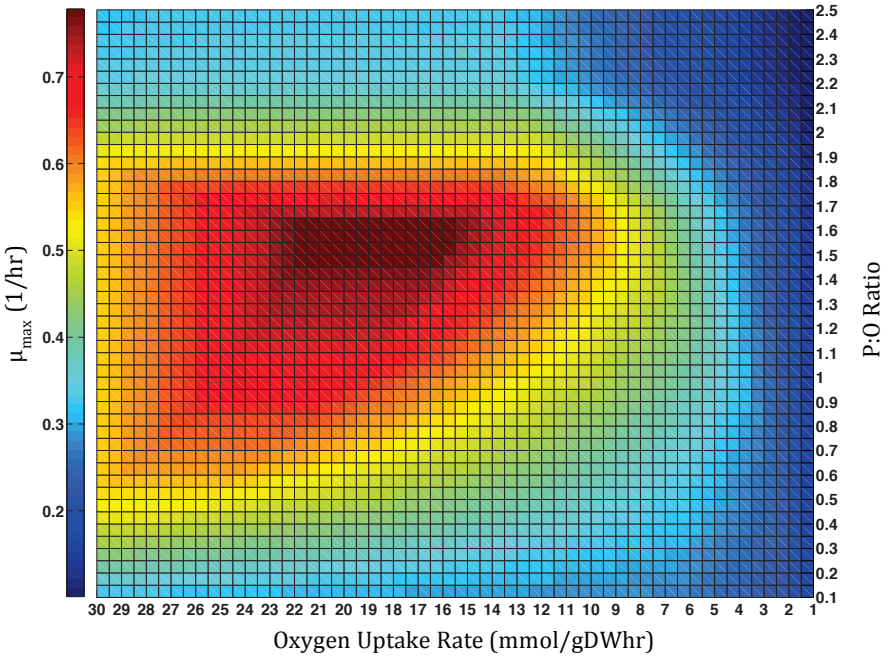

Figure 3-2: Feasible P:O range and corresponding specific growth rates for different OUR – Color-map

This approach enables us to perform sensitivity analysis by varying the P:O ratio over a broad range, and observe its effects on the theoretical maximum growth of *E. coli*. With open bounds on Oxygen Uptake Rate and a fixed glucose uptake rate, analysis with FBA and TFA indicated that the optimum P:O Ratio for *E. coli* is 1.575 (Figure 3.1 a-b). To investigate the effect of OUR on P:O ratio and specific growth rate, we repeated the same study by fixing different oxygen uptake rates. This *in silico* analysis of different oxygen uptake rates shows an optimum specific growth rate (0.78/hr, with glucose uptake rate of 8.16 mmol/gDWhr) at ~15.5-21.5 mmol/gDWhr with 1.575 as P:O ratio, which shows accordance with experimental optimum values (Figure 3.1). Moreover, when P:O ratio is forced to be bigger than 2.2, the cell cannot maintain high fluxes through cytochromes and ATP synthase, thus decreasing the ETC activity. If we force a high oxygen uptake flux in this condition, the cell uses this oxygen in other reactions,

such as MOX, malate oxidase. This signifies an important stoichiometric limit for *E. coli* in terms of P:O ratio.

After the optimum oxygen uptake rate of ~15.5 mmol/gDWhr, we do not observe an increase in the yield; on the contrary, if we force more oxygen than ~21.5 mmol/gDWhr, the specific growth rate of the cell decreases (See the plateau in Figure 3.1). *In silico* optimum P:O (1.3 to 1.575) values under different oxygen uptake rates are in the experimentally reported range, and it continuously increases till the optimum specific growth rate, and gets stabilized from this point, and stays around 1.55-1.6.

In silico and experimental P:O ratio is calculated by stoichiometric basis, which must be connected to the 'energy' term in bioenergetics. In order to reveal the relation between P:O ratio and bioenergetics of the cell in more details, we utilized a term, long known to have relation with the carbon yield of the cell[17], Gibbs free energy dissipation. Gibbs free energy dissipation is the amount of free energy that the cell releases to the environment. In this context, we used the Gibbs free energy dissipation term as the heat that must be released from the cell to produce 1 Carbon mole of biomass, which is the consensus term that represent the amount of biochemical work required to convert a carbon source into biomass and can be used to characterize chemotrophic microbial growth[17].

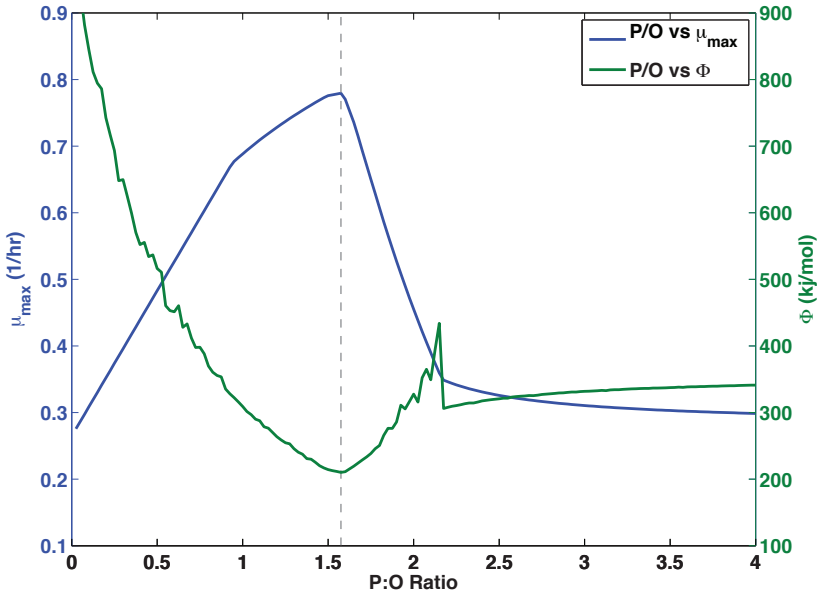

Figure 3-3: Gibbs Free Energy Dissipation vs. P:O Ratio and Growth

Firstly, we fixed the glucose uptake to the experimental value, and tested the behaviour of the cell under different P:O ratios, and calculated the corresponding growth and Gibbs free energy dissipation of the cell. We did not put any constraint on the oxygen uptake, to observe the direct relation between P:O ratio and dissipation. As we move along the P:O curve (Figure 3.3), we observe that minimum Gibbs free energy dissipation corresponds to the optimum P:O ratio and optimum growth. This suggests that the cell minimizes the amount of free energy that it disposes to the environment while it maximizes growth. However, the curve that represents the free energy dissipation is calculated based on 1 solution and is not tested for alternative solutions. Sampling methods can be used to study different flux profiles for FBA problems, however they cannot account for concentrations, since sampling cannot be performed for Mixed Integer Linear Programming problems like TFA. To study the extrema for free energy dissipation, we employed a new approach. The formulation of Gibbs free energy dissipation suggests a very important characteristic of the cells; the dissipation does not depend on the intracellular concentrations, since the contribution of the intracellular concentrations to Gibbs free energy term will cancel out when we sum up all the reaction $\Delta G'^{0}_{rxn,j}$

in the system. So the governing terms in the system are the standard Gibbs free energy of reactions, $\Delta G'^{0}_{rxn,j}$ and the extracellular metabolites (Materials and Methods for derivation). Moreover, by fixing the media composition concentrations from the literature, we reduce the flexibility that emerges from extracellular metabolites. In order to study this behaviour, in each P:O point along the curve, we calculated the maximum and maximum sum of standard transformed Gibbs free energy of all reactions (Figure 3.4), multiplied by the corresponding flux value. We call this term 'pseudo-dissipation'. As hypothesized, we observed that the 'real' free energy dissipation of the cell is accordant to its minimum and maximum pseudo-dissipation and are showing the same behaviour along the P:O Ratio curve. Highest growth coincides with the minimum pseudo free energy dissipation extrema, and the real dissipation that is calculated based on 1 solution shows the same behaviour. Another striking result is the smoothness of the curve for pseudo dissipation extrema; we do not observe any peaks and troughs. Moreover, the P:O ratios that have wider pseudo dissipation range are the ones that we observe the peaks and troughs in real dissipation based on 1 solution.

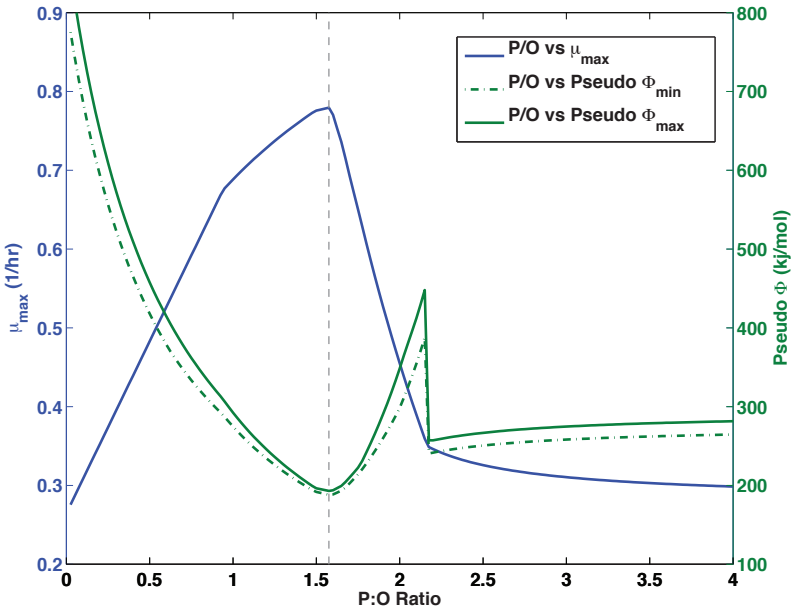

Figure 3-4**:** The relation between P:O ratio, specific growth (1/hr), and Gibbs free energy dissipation (kcal/mol) of aerobically grown *E. coli*. The highest grown, and optimum P:O ratio corresponds to the minimum Gibbs free energy dissipation, and real dissipation can vary between the minimum and maximum pseudo dissipation. The real dissipation value is based on 1 solution (within degrees of freedom of the system), hence it can fluctuate according to maximum and minimum pseudo dissipation. As the system moves towards optimum, the allowable range for dissipation shrinks. Vertical dashed line indicates the optimum P:O Ratio.

3.3.1 Affect of Oxygen Uptake Rate (OUR) on P:O Ratio and Dissipation Behaviour

Up to this point, in all simulations we only used an upper bound for OUR, which is 30 mmol/gDWhr. Since OUR is directly connected to P:O Ratio through Cytochrome enzymes, we decided to observe the system behaviour under different oxygenation states, and repeated the calculations with different OURs (Figure 3.5).

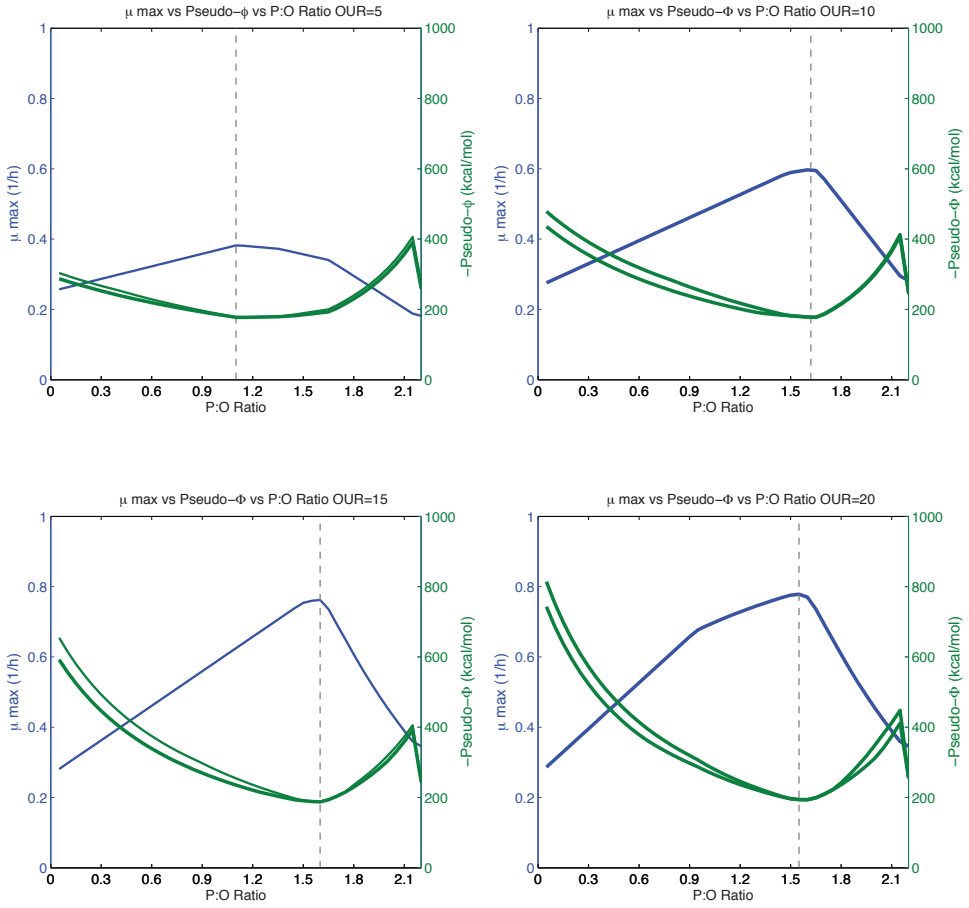

Figure 3-5: P:O vs. Oxygen Uptake Rate vs. Gibbs Free energy dissipation. This figure illustrates the relation between P:O ratio, specific growth rate and Gibbs free energy dissipation. Expectedly, Gibbs free energy dissipation is at minimum at the optimum specific growth rate for all 4 cases. The optimum P:O ratio differs for different OURs.

As stated before, under different oxygen availability, the optimum P:O Ratio will vary. For instance, at OUR=5mmol/gDWhr, the theoretical highest specific growth rate is achieved at P:O Ratio= 1.1. This is significantly smaller than the optimum under fully aerobic conditions, which is 1.575. However, the optimum P:O Ratio and minimum pseudo extrema still coincide and this behaviour is preserved for different OURs. This

shows that under different environmental conditions, the cell tries to minimize the Gibbs free energy dissipation.

However, still these data are not sufficient to explain the details on dissipation behaviour of *E. coli* since they represent the overall Gibbs free energy dissipation behaviour, and approaches the organisms as a black box with certain inputs and outputs. In order to investigate in more details, we studied on the Gibbs free energy dissipation characteristics in a reaction-by-reaction basis.

3.3.2 Analysis of Gibbs free energy dissipation of *E. coli* on reaction basis

In this chapter, we claimed that *E. coli, w*hile optimizing growth, minimizes the amount of free energy dissipated to the environment. *In silico* analysis gives the opportunity to determine the enzymatic reactions, or reaction sets that govern the dissipation behaviour of the organism under different P:O ratio and oxygen uptake rates. To achieve this goal, we performed an analysis on the GEM, by taking into account the following criteria: *i) keeping the cell under the optimized specific growth at each P:O point along the curve in Figure 3.2*. *ii) Keeping the cell under the pseudo dissipation range calculated from GEM. iii) Performing variability analysis on Gibbs free energy dissipation for each reaction.* By moving along the P:O curve, we determined all reactions that are having a significant contribution to free energy dissipation. Among these reactions, in each point along P:O curve, reactions, mainly from dehydrogenases, transhydrogenases, cytochrome oxidoreductases and quinone reductases (see Table 3.2) have the largest effect on the dissipation behaviour. Among these reactions, dehydrogenases and quinone reductases have the biggest effect (for full list Appendix Table A.1). Another conclusion from this analysis is that, even though we do not see every reactions active simultaneously along the P:O curve, dehydrogenases, transhydrogenases, cytochrome oxidoreductases and quinone reductases are interchangeably used by the cell (inside their own enzyme family), and the cell minimizes the amount of free energy dissipated to the environment by choosing the necessary, but optimum in terms of dissipation, biotransformations catalysed by these enzymes. In a recent study[99], the usage of cytochromes were reported to be directly related to the membrane occupancy, in this study, we claim that the usage

of the type of cytochromes does not depend only on the number of pumped protons and membrane occupancy, but also it is related to the amount of heat dissipated to the environment. Our final observation is that, as the ATP synthase enzyme synthesizes more ATP, it brings more dissipation burden to the cell; hence we claim that the ATP production of the cell is not only constrained by stoichiometric limitations, but also by thermodynamic properties of the organism under aerobic conditions.

Table 3-2: Enzymatic reactions that affect most the Gibbs free energy dissipation of aerobically grown *E. coli*. DIM is an abbreviation for DIMethyl.

Enzyme	Enzymatic Reaction Formula	$\Delta G'^{0}_{rxn}$
ATPM	ATP + H2O → ADP + H^+ + Phosphate	-7.76
ATPsynthase	ADP + Phosphate + 4H^+(p) → ATP + 3H^+(c) + H2O	-8.80
CYT_{bd1}	2H^+(c) + 0.5 O_2 + Ubiquinol8 → H2O + Ubiquinone8 + 2H^+(p)	-28.66
CYT_{bd2}	2H^+(c) + 0.5 O_2 + Menaquinol8 → H2O + Menaquinone8 + 2H^+(p)	-14.94
CYT_{bo3}	4H^+(c) + 0.5O_2 + Ubiquinol8 → H2O + Ubiquinone8 + 4H^+(p)	-20.38
GLUDy	L-Glutamate + H2O + $NADP^+$ → Oxoglutarate + H^+ + NADPH + NH4	11.23
MDH	L-Malate + NAD^+ → H^+ + NADH + 2-Oxaloacetate	4.75
MDH2	L-Malate + Ubiquinone8 → Oxaloacetate + Ubiquinol8	-11.22
NADH17pp	4H^+(c) + NADH + Menaquinone8 → NAD^+ + Menaquinol8 + 3H^+(p)	-13.13
NADH18pp	4H^+(c) + NADH + DIMmenaquinone8 → NAD^+ + DIMmenaquinol8 + 3H^+(p)	-2.28
SUCDi	Ubiquinone8 + Succinate → Fumarate + Ubiquinone8	-2.44
NADPHQR2	H^+ + NADPH + Ubiquinone8 → $NADP^+$ + Ubiquinol8	-15.58
NADPHQR3	H^+ + NADPH + Menaquinone8 → $NADP^+$ + Menaiquinol8	-28.30
NADPHQR4	H^+ + NADPH + Dimethylmenaquinone8 → $NADP^+$ + Dimethylmenaquinol8	-18.45
PDH	CoA + NAD^+ + Pyruvate → AcCoA + CO_2 + NAD^+	-10.53
THD_2	NADH + $NADP^+$ + 2H^+(p) → 2H^+(c) + NAD^+ + NADPH	-8.67

3.3.3 Relation between acetate production and Gibbs free energy dissipation of E. coli

The goal of the cell has been a debate for many years, and recently there are different approaches and hypothesis for the multiple optimality criteria for cells[100]. Until this point, we reported on optimally grown *E. coli* cells, without a significant by-product formation other than CO_2, under these conditions, the lowest Gibbs free energy dissipa-

tion and the optimal growth corresponded. To extend our study, we then studied the Gibbs free energy dissipation characteristics of by-product forming *E. coli.* Even under perfect conditions (aerobic, glucose, rich in salts media), *E. coli* is known to produce by-product acetate[101]. In order to test our hypothesis with acetate producing *E. coli*, firstly, we forced the experimental amount of acetate as by-product (4.95 mmol/gDWhr acetate from 8.16 mmol/gDWhr glucose) from the model, and tested P:O versus growth characteristics, which resulted in the same behaviour that we observed with theoretical maximum growth rate, with an optimum P:O value as 1.55. Also when we calculated the Gibbs free energy dissipation values to produce 1 Carbon mole of biomass, we saw that the optimum growth as 0.65/hr corresponds to the lowest Gibbs free energy dissipation along the P:O curve, which shows that the Gibbs free energy dissipation minimization still holds under by-product formation.

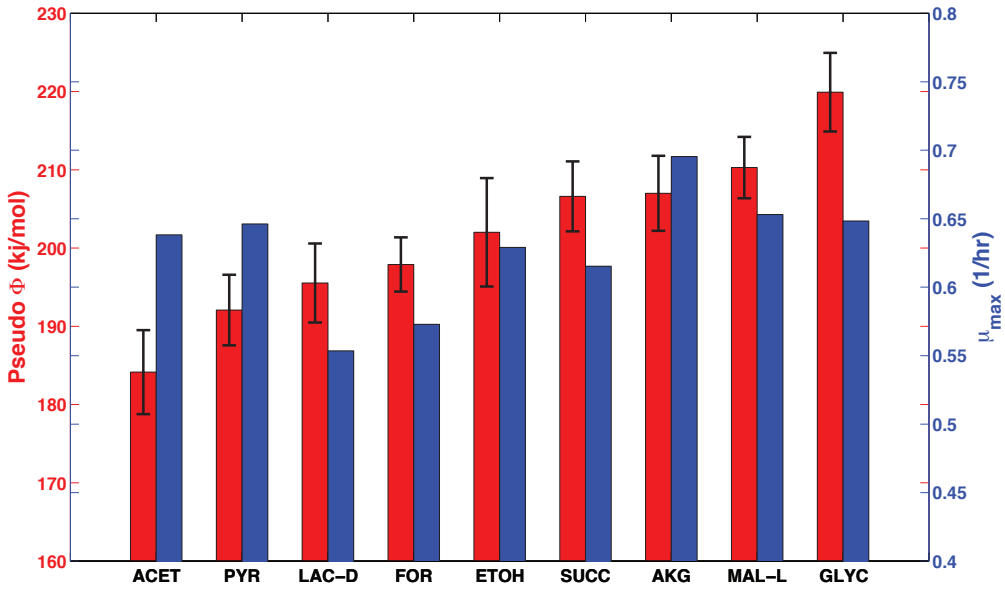

Figure 3-6**:** Gibbs free energy dissipation of the *E. coli* cells with different product formation: By forcing the same amount of carbon disposal from each by product, the corresponding Gibbs free energy dissipation (to form 1 Carbon mole of Biomass) and specific growth is calculated. Acetate disposal, even though it does not lead to highest specific growth rate, is observed to produce lowest dissipated free energy to the environment.

E. coli can produce different by-products under imperfect conditions, and it does a mixed acid fermentation under anaerobic conditions. Hence, a question regarding the production of acetate emerges. Why does the cell prefer to produce acetate rather than other by-products? In order to test the dissipation characteristics for each by-product disposal cases, we forced the same amount of carbon through each of them, while keeping CO_2 disposal unconstrained. We maximized the biomass production, and fixed the theoretical maximum growth rate and performed a variability analysis on the pseudo dissipation to observe the limits of free energy dissipation in each case. We observed that pseudo dissipation varies between 179kj/C-mol to 190 kj/C-mol of free energy, with a specific growth rate of 0.65/hr, which is very close to experimental growth rate of 0.61/hr in acetate producing case. All other 8 cases, namely, pyruvate, lactate, ethanol, oxoglutarate, succinate, malate, formate and glycerol, generated more free energy, even though 2 of them (pyruvate, oxoglutarate) resulted in higher specific growth rates (Figure 3.6). We claim that acetate is preferred over other by-products, since its disposal produces less heat than other metabolite secretions, or at least we can say Gibbs free energy dissipation is one criterion for this preference. Moreover, when acetate production pathway is blocked by knocking out 2 main genes, phosphotransacetylase and acetate kinase, the main by-product becomes pyruvate[102], which is the second optimal candidate according to the free energy dissipation analysis.

3.4 Conclusion

In this chapter, we focused on the bioenergetics efficiency of aerobic growth of *E. coli*. We have shown the consistency between *in silico* and experimental P:O ratio, as well as the growth characteristics and Gibbs free energy dissipation. We claim that the cell minimizes the amount of heat dissipated to the environment while growing with optimum capacity. We reported the enzymes that are responsible for determining the dissipation behaviour of the cell. We also showed the relation between the acetate producing *E. coli* cell and Gibbs free energy dissipation, and we hypothesize that the cell chooses acetate, because acetate production results in lowest Gibbs free energy dissipation even though it does not lead to the highest growth. Studying on the bioenergetics efficiency of the cell can be further used to determine the limitations of the cell for growth, or for pro-

duction of a certain industrial chemical production. We suggest that Gibbs free energy dissipation can be used as a selection criterion for the novel pathways for the production of target chemicals in host cell.

Chapter 4 – *lumpGEM*: Systematic Generation of Biosynthetic Subnetworks for Biomass Building Blocks

4.1 Introduction

Stoichiometric models have been extensively used since 1980s[103-105] and prediction capabilities of these networks have been proven to be very useful. The size and structure of these models varied among different studies. One of the pioneering studies on *E. coli* through a small stoichiometric model is performed by Varma et al. [106,107] in where the authors described the model as composed of core carbon metabolism pathways namely, glycolysis, pentose phosphate pathway, TCA cycle and formation of some by-product formations accompanied by a part of the Electron Transport Chains (ETC). This stoichiometric definition is further extended by the integration of a biomass composition formulation that is provided in the classic text published by Neidhardt[108]. In this textbook, *E. coli* metabolism has been explained extensively, and all the components (amino acids, lipids, DNA, RNA etc.) that constitute 1gDW of cell were reported based on previous experiments[21]. Moreover, the amounts of 12 precursor metabolites from the core carbon metabolism (erythrose-4-phosphate, ribose-5-phosphate, pyruvate, alpha-ketoglutarate, phosphoenolpyruvate etc.) along with the requirement of cofactors (ATP, NADH, NADPH) and inorganic compounds (S, NH4) to synthesize these biomass building blocks (BBB) were estimated. This information allowed the authors to develop a model that can mimic the growth behaviour of the organism without including the complex biosynthesis routes for each individual biomass building block. With such a small stoichiometric representation of the core metabolism (~50 reactions), authors were

able predict many aspects of *E. coli* physiology. Similar metabolic models have been used in many other studies[109-111].

In the following years, with the development of sequencing and high-throughput technologies, the gene-protein-reaction (GPR) associations[112] have been improved and this sparked off the number of sequenced genomes[113,114]. This accumulation of knowledge eventually has led to the development of Genome Scale Metabolic Networks (GEMs), firstly for *E. coli* [115], which encapsulate all the known biochemistry of organisms. These comprehensive representations of metabolism are accompanied by biomass formulations that account for 1gDW of cell composition[10]. The contribution of each biomass building block is either determined empirically, or approximated from phylogenetically close species[116].

The emerging GEM era has produced many metabolic models for various organisms [30,117,118] and their strains[95,119] and proved to be extremely useful for many different purposes from strain design for biosynthesis of industrial chemicals to drug discovery[120]. The applicability of the GPR method eventually has led to the development of GEMs for higher organisms such as mouse[121], plants[76] and human[122]. Although GEMs are widely used and they have provided important insight and guidance, the charm of small and yet predictive models still shines the eyes of researches for many purposes. And while the biomass formulation of Neidhardt is still in use and has proven to be valid in the last 25 years, as Pramanik et al. [111] have shown, the changes in the biomass composition has significant effects on the internal fluxes, thus should be considered very carefully. In this respect, the extended and curated biochemistry in GEMs can be used to validate and to improve the approximations made by Neidhardt, and to extend it for every biomass building block defined in biomass compositions of GEMs and to account for alternative synthesis routes.

Towards this, we developed lumpGEM, an algorithm that identifies all the alternative reaction subnetworks that should be used to produce a cellular metabolite or biomass building block from defined subset of metabolites, such as a core metabolic network/subsystem. In this study, we focused on the biosynthesis pathways of biomass

building blocks of *E. coli* IJO1366 [119], and with our tool **lumpGEM**, we have identified known and possible other alternative synthesis routes/subnetworks for all BBBs from *core* carbon metabolism as defined in[106,107]. We demonstrated that lumpGEM is capable of building lumped reactions in where the contribution of each *core* carbon metabolite to synthesize a biomass building block is identified and properly accounted.

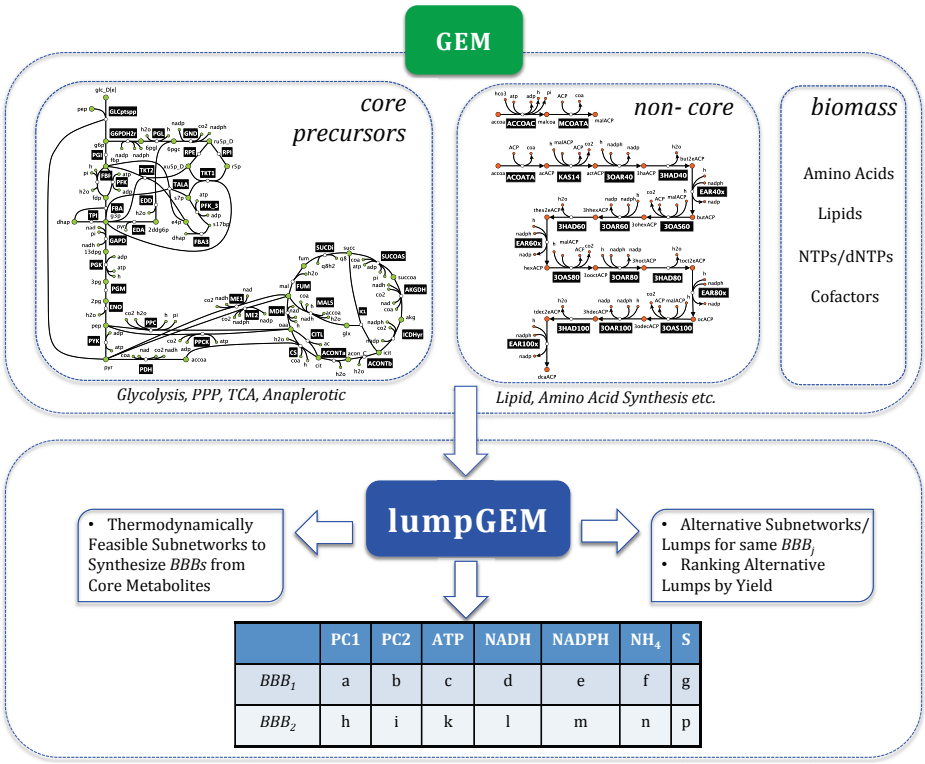

Figure 4-1: Inputs and outputs for lumpGEM. By defining the *core* precursors (PC1, PC2, …), biomass building blocks (*BBBs*), and *non-core* parts of metabolism, the GEM is provided to lumpGEM. The output of lumpGEM is thermodynamically feasible subnetworks, which with the *core*, is capable of ing *BBBs*. The MILP characteristic of lumpGEM allows the building of alternative subnetworks and lump reactions for the same BBB_j, and it ranks them according to yield.

Using these lumped reactions, we performed a comparison between the approximations of Neidhardt and the values generated by lumpGEM. Such studies will help us to understand the capabilities of *E. coli* 'per' biomass building block and identify the flexibility of the organism to survive by activating different parts of the metabolism to accumulate biomass. The generality of the method makes it applicable to any GEM that has a well-

defined biomass composition. In addition, lumpGEM can generate lumped reactions from any part of the metabolism for any target metabolite, either a biomass building block or a biochemical and chemical compound or sets of compounds, which makes it versatile to be used for different purposes.

4.2 Materials and Methods

4.2.1 Preliminary Definitions

BBB: Biomass building block.

M^{core}: Metabolites that belong to core system, in this specific case, defined in Varma et al. [106,107].

R^{core}: Reactions that belong to core system, in this specific case, defined in Varma et al. [106,107].

M^{GEM}: Metabolites that belong to GEM that do not belong to M^{core}.

R^{GEM}: Reactions that belong to GEM that do not belong to R^{core}.

S^j: Subnetwork (set of reactions) that synthesizes BBB_j other than M^{core} R^{core} and composed of M^{GEM} and R^{GEM}.

M^{sub}: Metabolites that belong to S^j.

R^{sub}: Reactions that belong to S^j.

FUSE: Binary decision variable that controls the flux through each R^{GEM}. When decision is 0, the reaction is active.

4.2.2 Generating Subnetworks for each BBB

a. Decompose the biomass composition of GEM to each of its components, such as alanine, tyrosine, biotin, etc. Built a new GEM model by allowing the individual production of each *BBB*.
b. Define M^{core} and R^{core}.
c. Split all the reactions in GEM in Step a. into forward $F_{rxn,i}$ and backward $B_{rxn,i}$ components.
d. Create binary use variables $FUSE_i$ for each R_i^{GEM}
e. Generate a constraint for each R^{GEM} that will control the flux through these reactions as:

$$F_{rxn,i} + B_{rxn,i} + C.FUSE_i \leq C$$

C is the number of carbon atoms that the cell uptakes from its surrounding. If the cell can uptake multiple carbon sources, and the number of carbon atoms is not definite, an arbitrary big number can substitute for *C*.

Postulate 1: Binary control is unbiased to reaction directionality. This means that R_i^{GEM} that is controlled by $FUSE_i$ can operate in both directions if the existing constraints (mass balance and thermodynamics) allow it.

f. Apply thermodynamics constraints for M^{core} and R^{core} as defined in[13,14].
g. Build the following MILP formulation for each BBB_j:

Maximize

$$\sum_{i}^{\#\ of\ R^{GEM}} FUSE_i$$

such that:

$$S.v = 0 \qquad (4.1)$$

$$v_{BBB,j} \geq n_{j,GEM} \cdot \mu_{max} \quad (4.2)$$

$v_{BBB,j}$: The sink that is created in Step 1.a for BBB_j for its biosynthesis.

μ_{max}: Theoretical maximum specific growth rate for the given physiology in 1/hr units.

$n_{j,GEM}$: The stoichiometric coefficient for BBB_j in mmol/gDW unit as defined in original GEM.

Postulate 2: Any M^{core} is a potential precursor for the biosynthesis of BBB_j.

Postulate 3: Maximizing for the sum of $FUSE_i$ results in the smallest subnetwork S^j to produce BBB_j from M^{core}. This subnetwork is not necessarily composed of only linear pathways as reported in databases such as KEGG[22], SEED[123] or EcoCyc[124] etc. and may include branches.

Postulate 4: The flux distribution for each generated subnetwork cannot guarantee an optimum flux distribution that will specify the individual stoichiometric contribution of each M^{core} to synthesize BBB_j due to the degrees of freedom (DOF) that the system has.

Moreover, the GEM built in Step f. for generating subnetworks is partially constraint with thermodynamics, and these subnetworks are constrained by only mass-balance.

To overcome these limitations, we have built the following MILP formulation for each S^j:

a. Generate a model comprised of:
    i. M^{core} and R^{core}
    ii. S^j
    iii. v_{BBB}
b. Apply thermodynamic constraints on this model as described in [13,14].
c. Build the following formulation for each S^j:

Minimize

$$\sum_{k}^{\#\ of\ R^{sub}} F_{R_k^{sub}} + \sum_{k}^{\#\ of\ R^{sub}} R_{R_k^{sub}}$$

such that

$$S.v = 0$$

$$v_{BBB,j} \geq 1 \tag{4.3}$$

$F_{R_k^{sub}}$ and $R_{R_k^{sub}}$ are integer variables.

Postulate 5: Minimizing the sum of net flux through R^{sub} generates a stoichiometrically proportional flux distribution in the subnetwork S^j. This leads to the exact stoichiometric expenditure of each M^{core} to synthesize BBB_j.

d. Lump R^{sub} with respect to the flux distribution obtained above, Step c. This is collapsing the reactions into 1 overall reaction that is stoichiometrically equivalent to the flux distribution generated above.

Generating Alternative Subnetworks for each BBB_j

To identify alternative subnetworks for BBB_j, GEM is further constrained with the following constraint after generating each S^j with an iterative manner[125].

$$\sum_{k}^{\#\ of\ R^{sub}} FUSE_{R_k^{sub}} > 0$$

Postulate 6: Since R_k^{sub} is active if only $FUSE_{R_k^{sub}} = 0$, the next solution will have at least 1 different reaction from the previous solution. Aftermath, the same procedure is applied for the newly generated $S^{j,2}$.

4.3 Results and Discussion

When the genome scale model is optimized for maximum specific growth with Flux Balance Analysis, the contribution of M^{core} (the metabolites defined in Varma stoichiometry) to synthesize a biomass building block is not evident from the flux distribution due to the degrees of freedom that the system has, and the alternative routes that a BBB can be synthesized. In order to overcome this limitation, we have developed **lumpGEM**, a Mixed-Integer Linear Programming (MILP) based tool (See Material and Methods) to reveal the contribution of M^{core} and R^{core} (reactions defined in Varma stoichiometry) in an optimum manner. MILP formulations have been often used on biochemical networks for many purposes[126-129] since they allow the control of reactions with an on/off manner. We made use of this binary decision in order to control the flux through the reactions of R^{GEM}(the reactions defined in GEM network other than R^{core}), and lumpGEM allowed us to build minimal subnetworks that can synthesize BBBs from any defined part of the metabolism in GEMs, in this specific case, the core carbon metabolism.

The biomass formulation defined in *E. coli* IOJ1366 is very well characterized and detailed and contains 102 biomass building blocks. It is mainly composed of amino acids, lipids, nucleoside triphosphates (NTPs), deoxy-NTPs and inorganic compounds (Nickel, Zinc, Iron, etc) along with cofactors such as NAD^+/NADH, $NADP^+$/NADPH, CoA/AcCoA, and FAD. Experimental estimates of the growth associated ATP maintenance and production of diphosphate are also included in the biomass composition.

The main difference between our approach for generating synthesis routes for the BBBs and the database-based analysis is that the *subnetworks* that our method generates may include branches from linear synthesis pathways. The difference emerges from the mass conservation constraint that we force during our analysis. For instance, the smallest subnetwork that lumpGEM generated for the synthesis of histidine is composed of 21 reactions and the precursors are ribose-5-phosphate (R5P) and oxaloacetate. In the databases, the linear pathway for histidine synthesis is composed of 10 steps and has ribose-5-phosphate as the only precursor. When we analyse the 21-reaction subnetwork, we see branching points in the linear route from R5P to histidine. Due to the mass

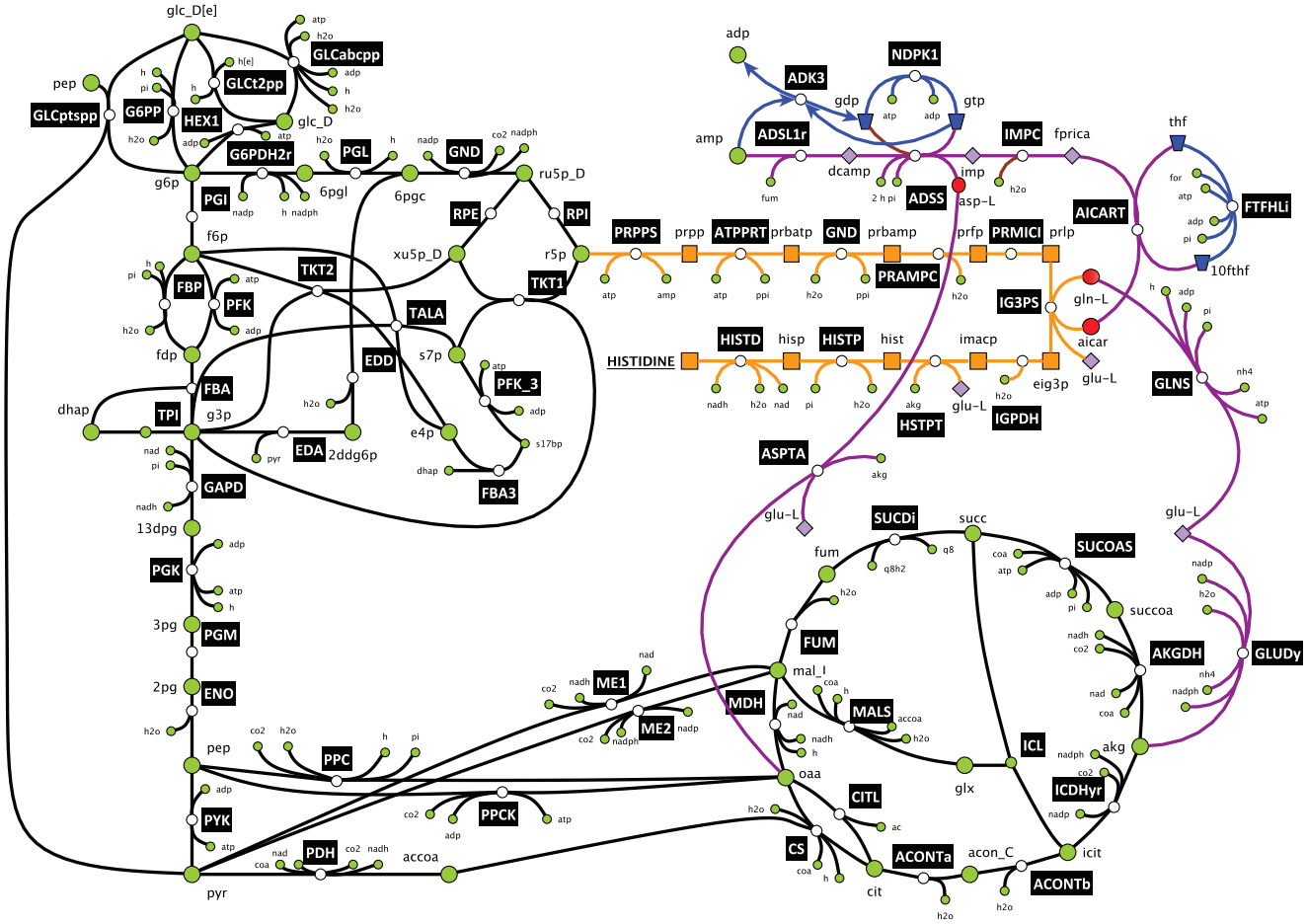

Figure 4-2: Synthesis of histidine from ribose-5-phosphate. The orange reactions are part of the linear synthesis route in the databases. The red, blue and purple metabolites are the *non-core* metabolites in the subnetwork. Red metabolites appear as the *non-core* metabolites in the database pathways. Finally, blue metabolites appear as the unbalanced metabolites along the purple reactions, and the blue reactions are balancing them. The subnetwork cannot produce any histidine from *core* network without including the purple reactions due to mass balance constraints.

balance constraint, three metabolites, 1-(5-Phosphoribosyl)-5-amino-4-imidazolecarboxamide, L-Glutamine and diphosphate (metabolites colored with red in IG3PS and PRATPP reactions in Figure 4.2) cannot be balanced in a network that is composed of *core* reactions and the linear pathway from ribose-5-phophate to histidine. Hence, the generated sets of reactions are not only the linear routes from precursor metabolites to biomass building blocks, but balanced subnetworks with stoichiometrically proportional branches. lumpGEM captured reactions from different subsystems to produce histidine, alanine and aspartate metabolism, anaplerotic reactions, folate metabolism, glutamate metabolism, histidine metabolism, nucleotide salvage pathway, purine and pyrimidine biosynthesis subsystems. In addition, the lumped reaction that is generated from this subnetwork (see Materials and Methods) has only *core* metabolites and biomass building blocks on both reactants and products sides. This representation is similar to Neidhardt's definition, since he also described the stoichiometric expenditure of *core* metabolites in his estimations. Similar to our analysis, the values that Neidhardt et al. reported for the synthesis of histidine are different than the linear route that is reported in databases. This gives the impression that Neidhardt et al. also took into account the expenditure of the *non-core* metabolites along the linear pathway. However, he reported only 1 value for each biomass building block. What lumpGEM allows us to perform is to build alternative subnetworks and possible lumped reactions for the same BBB_j. In this specific case, with the minimum S^j_{min} size as 21, lumpGEM generated 12 subnetworks, and 3 unique lumped reactions. *This signifies that the overall lumped reactions of different subnetworks can be the same.* This has been observed also in a previous study by Henry et al. that focuses on pathway generations for an industrial chemical[63].

Table 4-1: Expenditure of Cofactors and Precursor metabolites for Histidine Synthesis and the comparison between values reported by Neidhardt and the minimum-sized subnetworks' lumps by lumpGEM. 1-C in lumpGEM represents Formate.

HISTIDINE	Precursor	ATP	NADH	NADPH	1-C	NH4	S
Neidhardt	1 R5P	6	-3	1	1	3	0
lumpGEM-1	1 R5P, 1 OAA	5	-2	2	1	3	0
lumpGEM-2	1 R5P, 1 OAA	7	-2	2	1	3	0
lumpGEM-3	1 R5P, 1 OAA	9	-2	2	1	3	0

Oxaloacetate appears as a precursor to balance the *non-core* metabolite glutamine in the subnetwork, which is not reported in Neidhardt precursors. With a few exceptions like this, the approximations by Neidhardt for the consumption of cofactors and inorganics are very close to the values that lumpGEM reports (Table 4.1). These results explain why these approximations could be used in metabolic networks and how these networks could predict many aspects of different physiology.

The differences between the alternative subnetworks $S^j_{min,1}$ $S^j_{min,2}$ … may emerge from different reactions in the 'linear' pathway from main precursor to the biomass building blocks, or from the other *non-core* reactions which are balancing the *non-core* metabolites in the linear route. These two sources of differences, and especially the latter, may result in an explosion in the number of subnetworks that can be generated for some of the biomass building block. For metabolites like amino acids, which are not so far from the core carbon metabolism, the number of smallest subnetworks S^j_{min} are not many (Table 4.2), however for big molecules, such as lipids, there exists hundreds of alternative routes (Table 4.3, full list for all *BBBs* Appendix Table A.2). For example, Phosphatidylglycerol (dihexadecanoyl, n-c16:0) is a lipid with a S^j_{min} of 40 reactions, and has 256 alternative subnetworks with 16 unique lumped reactions. In the first subnetwork generated by lumpGEM, within the 40 reactions, 34 of them are part of linear synthesis route and 6 of them are balancing *non-core* metabolites. In all lipid cases, hundreds of different subnetworks are represented by tens of unique lumped reactions (Table 4.3).

Table 4-2: Lumped Reactions and Statistics for Amino Acids. The numbers (i:j:k) represent *i,* size of subnetworks for amino acids (2:1:1), *j,* the number of alternative subnetworks (2:1:1) and k, the number of unique lumped reactions (2:1:1), respectively. The rows under the lumped reactions represent the values reported by Neidhardt. In this table the size corresponds only to smallest size subnetworks (S^{j}_{min}) for each BBB.

BIOMASS BUILDING BLOCK	LUMPED REACTIONS
L-ALANINE	2:01:01 H + NADPH + NH4 + PYR -> ALA-L + H2O + NADP Neidhardt: PYR + NADPH + NH4 -> ALA-L
L-ARGININE	13:02:02 ACCoA + AKG + **3 ATP** + CO2 + 4 NADPH + 4 NH4 + OAA -> AC + **3 ADP** + ARG-L + CoA + FUM + **H** + **2 H2O** + 4 NADP + **3 PI** ACCoA + AKG + **4 ATP** + CO2 + 4 NADPH + 4 NH4 + OAA -> AC + **4 ADP** + ARG-L + CoA + FUM + **2 H** + **H2O** + 4 NADP + **4 PI** Neidhardt: AKG + 4 ATP + 4 NADPH + 4 NH4 -> NADH + ARG-L
L-ASPARAGINE	5:02:02 **2 ATP** + NADPH + 2 NH4 + OAA -> **2 ADP** + ASN-L + H + NADP + **2 PI** **ATP** + NADPH + 2 NH4 + OAA -> **ADP** + ASN-L + H2O + NADP + **PI** Neidhardt: 4 ATP + 2 NH4 + OAA -> ASN-L
L-ASPARTATE	2:01:01 H + NADPH + NH4 + OAA -> ASP-L + H2O + NADP Neidhardt: NADPH + NH4 + OAA -> ASP-L
L-CYSTEINE	16:06:02 3PG + ACCoA + **4 ATP** + NAD + 5 NADPH + NH4 + SO4 -> AC + **4 ADP** + CoA + CYS-L + NADH + 5 NADP + **5 PI** 3PG + ACCoA + **3 ATP** + H + NAD + 5 NADPH + NH4 + SO4 -> AC + **3 ADP** + CoA + CYS-L + **H2O** + NADH + 5 NADP + **4 PI** Neidhardt: 3PG + 4 ATP + 5 NADPH + NH4 -> CYS-L
L-GLUTAMINE	2:02:02 AKG + **ATP** + NADPH + 2 NH4 -> **ADP** + GLN-L + **H2O** + NADP + **PI** AKG + **2 ATP** + NADPH + 2 NH4 -> **2 ADP** + GLN-L + **H** + NADP + **2 PI** Neidhardt: AKG + NADPH + NH4 -> GLN-L
	1:01:01

L-GLUTAMATE	AKG + H + NADPH + NH4 -> GLU-L + H2O + NADP Neidhardt: AKG + NADPH + NH4 -> GLN-L
GLYCINE	8:02:02 **3PG** + **H2O** + **NAD** + NH4 -> **FOR** + GLY + **3 H** + **NADH** + **PI** **2 ATP** + **2 H** + **3 NADPH** + NH4 + **OAA** -> **ACALD** + **2 ADP** + GLY + **3 NADP** + **2 PI** Neidhardt: 3PG + NADPH + NH4 -> GLY + NAD + 1-C
L-HISTIDINE	21:12:03 **5 ATP** + FOR + 2 NAD + 2 NADPH + 3 NH4 + OAA + R5P -> **5 ADP** + FUM + **8 H** + HIS-L + 2 NADH + 2 NADP + **6 PI** **7 ATP** + FOR + **2 H2O** + 2 NAD + 2 NADPH + 3 NH4 + OAA + R5P -> **7 ADP** + FUM + **10 H** + HIS-L + 2 NADH + 2 NADP + **8 PI** **9 ATP** + FOR + **4 H2O** + 2 NAD + 2 NADPH + 3 NH4 + OAA + R5P -> **9 ADP** + FUM + **12 H** + HIS-L + 2 NADH + 2 NADP + **10 PI** Neidhardt: 6 ATP + 1-C + NADPH + 3 NH4 -> HIS-L + 3 NAD
L-ISOLEUCINE	12:01:01 2 ATP + 5 H + 5 NADPH + NH4 + OAA + PYR -> 2 ADP + CO2 + 2 H2O + ILE-L + 5 NADP + 2 PI Neidhardt: 2 ATP + 5 NADPH + OAA + PYR + NH4 -> ILE-L
L-LEUCINE	10:01:01 ACCoA + 2 H + NAD + 2 NADPH + NH4 + 2 PYR -> 2 CO2 + CoA + H2O + LEU-L + NADH + 2 NADP Neidhardt: ACCoA + 2NADPH + NH4 + 2 PYR -> LEU-L + NADH
L-LYSINE	11:01:01 ATP + 4 H + 4 NADPH + 2 NH4 + OAA + PYR + SUCCoA -> ADP + CO2 + CoA + 2 H2O + LYS-L + 4 NADP + PI + SUCC Neidhardt: 2 ATP + 4 NADPH + 2 NH4 + OAA + PYR -> LYS-L
L-METHIONINE	25:06:02 2 3PG + ACCoA + **5 ATP** + **2 H** + NAD + 9 NADPH + 2 NH4 + OAA + SO4 + SUCCoA -> AC + **5 ADP** + 2 CoA + GLY + **H2O** + MET-L + NADH + 9 NADP + **7 PI** + PYR + SUCC 2 3PG + ACCoA + **4 ATP** + **3 H** + NAD + 9 NADPH + 2 NH4 + OAA + SO4 + SUCCoA -> AC + **4 ADP** + 2 CoA + GLY + **2 H2O** + MET-L + NADH + 9 NADP + **6 PI** + PYR + SUCC Neidhardt: 7 ATP + 8 NADPH + NH4 + 1-S + 1-C -> MET-L
L-PHENYLALANINE	11:01:01 ATP + E4P + 2 NADPH + NH4 + 2 PEP -> ADP + CO2 + H + 2 H2O + 2 NADP + PHE-L + 4 PI Neidhardt: E4P + 2 NADPH _ NH4 + 2 PEP -> PHE-L

L-PROLINE	5:01:01 AKG + ATP + 2 H + 3 NADPH + NH4 -> ADP + 2 H2O + 3 NADP + PI + PRO-L Neidhardt: ATP + AKG + 3 NADPH + NH4 -> PRO-L
L-SERINE	4:01:01 3PG + NAD + NADPH + NH4 -> H + NADH + NADP + PI + SER-L Neidhardt: 3PG + NADPH + NH4 -> NADH + SER-L
L-THREONINE	7:01:01 2 ATP + 2 H + 3 NADPH + NH4 + OAA -> 2 ADP + 3 NADP + 2 PI + THR-L Neidhardt: 2 ATP + 3 NADPH + NH4 + OAA -> THR-L
L-TRYPTOPHAN	17:02:02 **4 ATP** + E4P + NADPH + 2 NH4 + 2 PEP + R5P -> **4 ADP** + CO2 + G3P + **6 H** + **H2O** + NADP + 7 PI + TRP-L **3 ATP** + E4P + NADPH + 2 NH4 + 2 PEP + R5P -> **3 ADP** + CO2 + G3P + **5 H** + **2 H2O** + NADP + 6 PI + TRP-L Neidhardt: 5 ATP + E4P + 3 NADPH + 2 NH4 + 2PEP + R5P -> 2 NADH + TRP-L
L-TYROSINE	11:01:01 ATP + E4P + NAD + 2 NADPH + NH4 + 2 PEP -> ADP + CO2 + 2 H + H2O + NADH + 2 NADP + 4 PI + TYR-L Neidhardt: ATP + E4P + 2 NADPH + 2 PEP -> NADH + TYR-L
L-VALINE	5:01:01 3 H + 2 NADPH + NH4 + 2 PYR -> CO2 + 2 H2O + 2 NADP + VAL-L Neidhardt: 2 NADPH +NH4 + 2 PYR -> VAL-L

The small number of alternative subnetworks for amino acids also shows that the number of *non-core* metabolites that appeared along the linear synthesis route is small, since the main explosion in the number of subnetworks emerges from these reactions. As an example, all 12 alternative subnetworks for histidine include the linear 10 steps route from R5P to histidine and alternative subnetworks are generated by other *non-core* reactions. Moreover, there is a clear correlation between the number of alternative subnetworks and the size of the minimal subnetworks. Most of the amino acids that have more that 2 subnetworks require more than 10 steps for their biosynthesis.

Table 4-3: The size of smallest subnetworks for some lipids, the number of alternative subnetworks for each of them and the corresponding number of lumped reactions.

BIOMASS BULDING BLOCK	SIZE OF SUBNETWORK	# OF SUBNEWORKS	# OF LUMPED REACTIONS
PHOSPHATIDYLETHANOLAMINE (DIHEXADECANOYL, N-C16:0)	44	245	16
PHOSPHATIDYLETHANOLAMINE (DIHEXADEC-9ENOYL, N-C16:1)	44	109	14
PHOSPHATIDYLETHANOLAMINE (DIOCTADEC-11-ENOYL, N-C18:1)	48	256	16
PHOSPHATIDYLGLYCEROL (DIHEXADECANOYL, N-C16:0)	40	256	16
PHOSPHATIDYLGLYCEROL (DIHEXADEC-9-ENOYL, N-C16:1)	40	127	14
PHOSPHATIDYLGLYCEROL (DIOCTADEC-11-ENOYL, N-C18:1)	44	242	16
CARDIOLIPIN (TETRAHEXADECANOYL, N-C16:0)	41	512	32
CARDIOLIPIN (TETRAHEXADEC-9-ENOYL, N-C16:1)	41	255	28
CORE OLIGOSACCHARIDE LIPID A	75	128	24

4.3.1 Complex Biomass Components and Biomass Associated Processes

Apart from the inorganics, the biomass formulation has 87 metabolites on the reactant side, and lumpGEM generated subnetworks and lump reactions for 86 of them. The only metabolite that we did not perform lumping is ATP, which is a part of an ATP hydrolysis (Growth Associated Maintenance - GAM) reaction in the biomass formulation. The hy-

drolysis reaction of biomass is not stoichiometrically balanced, and has disproportional coefficients for the participating metabolites. When the Varma network is tested for the ability to hydrolyse ATP with the coefficients set in the GEM biomass formulation, it fails. However, if the hydrolysis reaction is balanced with the stoichiometric coefficient of ATP (54.12 mmol/gDW in this case), Varma network is able to hydrolyse this amount. In this case, ATP/ADP pool becomes a moiety and does not require any *de novo* synthesis of any metabolite (ATP or ADP) from the carbon source. Since the GEM is capable of accumulating biomass, it signifies that there are *non-core* reactions necessarily active to direct the carbon flow through biomass for ATP hydrolysis. In order to identify the subnetwork(s) for growth-associated maintenance (GAM), we followed the same procedure that we did for biomass building blocks, and we built a GEM with an additional GAM reaction with the coefficients of hydrolysis metabolites in GEM's biomass. Then, by forcing a flux of 1 mmol/gDWhr through this reaction, and by minimizing the number of *non-core* reactions, we generated minimal subnetworks that allow this hydrolysis. The resulting networks are composed of 27 reactions with 24 alternatives. The number of unique lumped reactions is 8. When we analyse the lumped reactions individually, we observe that these subnetworks are synthesizing *de novo* ADP to compensate for the smaller ADP coefficient compared to ATP (53.95 mmol/gDW compared to 54.12 mmol/gDW) in the biomass formulation.

With the lumped reactions generated for GAM, the only metabolite in the biomass formulation that we did not perform any analysis on is diphosphate. Diphosphate is secreted by biomass formulation; so there is not a synthesis pathway for this metabolite. By following the same procedure that we have followed for GAM, we generated *subnetworks* for the secretion of diphosphate. The size of S_{min}^{j} is 1, and there are 2 alternative reactions in R^{GEM} that can substitute each other for diphosphate secretion: Inorganic Diphosphatase (PPA) and Polyphosphate Kinase (PPKr). Both of these reactions are decomposing the secreted diphosphate into phosphate and proton, the former with water and the latter with ATP/ADP cofactor pair.

4.3.2 Ranking Alternative Lumped Reactions – Yield Analysis

When we analyse the alternative lumped reactions for the same biomass building block, we see different requirement of precursors, cofactors, nitrogen and sulphur. As explained before, this is an expected behaviour, and a detailed analysis could also suggest which lumped reaction is more suitable for specific studies. One of the main criteria to rank the lumped reactions is their capability to synthesize the BBB_j from the carbon source, and specifically the yield of *BBB* on the carbon source. In order to calculate the yield per lumped reaction, we built a 'mini' core model for each of them, which is composed of M^{core} - $R^{core,}$ V_{BBB} and the lumped reactions under study. By optimizing the synthesis of the BBB_j and calculating the C-mole yield over the carbon source of interest, specifically glucose, we ranked the alternative lumped reactions for each BBB_j. Interestingly, different lumped reactions can produce different amounts of biomass building blocks over a wide range of yield amounts (Table 4.4).

Table 4-4: The lumped reactions generated for deoxynucleoside triphosphate dTTP. The lumped reactions are sorted based on their *carbon mole dTTP synthesis* / *carbon-mole glucose uptake* yield.

BBB	LUMPED REACTIONS	Yield
dTTP	6 ATP + FOR + H + 4 NADPH + 2 NH4 + OAA + Q8 + R5P <=> 6 ADP + dTTP + 3 H2O + 4 NADP + 4 PI + Q8H2	0.83
	6 ATP + FOR + H + MQN8 + 4 NADPH + 2 NH4 + OAA + R5P <=> 6 ADP + dTTP + 3 H2O + MQL8 + 4 NADP + 4 PI	0.83
	8 ATP + FOR + 4 NADPH + 2 NH4 + OAA + Q8 + R5P <=> 8 ADP + dTTP + H + H2O + 4 NADP + 6 PI + Q8H2	0.83
	8 ATP + FOR + MQN8 + 4 NADPH + 2 NH4 + OAA + R5P <=> 8 ADP + dTTP + H + H2O + MQL8 + 4 NADP + 6 PI	0.83
	6 ATP + CoA + FOR + H + 3 NADPH + 2 NH4 + OAA + PYR + Q8 + R5P <=> ACCoA + 6 ADP + CO2 + dTTP + 3 H2O + 3 NADP + 4 PI + Q8H2	0.71
	6 ATP + CoA + FOR + H + MQN8 + 3 NADPH + 2 NH4 + OAA + PYR + R5P <=> ACCoA + 6 ADP + CO2 + dTTP + 3 H2O + MQL8 + 3 NADP + 4 PI	0.71
	8 ATP + CoA + FOR + 3 NADPH + 2 NH4 + OAA + PYR + Q8 + R5P <=> ACCoA + 8 ADP + CO2 + dTTP + H + H2O + 3 NADP + 6 PI + Q8H2	0.71
	8 ATP + CoA + FOR + MQN8 + 3 NADPH + 2 NH4 + OAA + PYR + R5P <=>	0.71

	ACCoA + 8 ADP + CO2 + dTTP + H + H2O + MQL8 + 3 NADP + 6 PI	
	6 ATP + FOR + FUM + H + 4 NADPH + 2 NH4 + OAA + R5P <=> 6 ADP + dTTP + 3 H2O + 4 NADP + 4 PI + SUCC	0.63
	8 ATP + FOR + FUM + 4 NADPH + 2 NH4 + OAA + R5P <=> 8 ADP + dTTP + H + H2O + 4 NADP + 6 PI + SUCC	0.63
	8 ATP + CoA + FOR + FUM + 3 NADPH + 2 NH4 + OAA + PYR + R5P <=> ACCoA + 8 ADP + CO2 + dTTP + H + H2O + 3 NADP + 6 PI + SUCC	0.56

dTTP is a deoxy nucleoside triphosphates and its main precursor for all the generated subnetworks is ribose-5-phosphate (R5P) from Pentose Phosphate Pathway. Consequently, the *core* network can supply the same amount of carbon to all the generated subnetworks and corresponding lumped reactions. One explanation for the differences in yields (Table 4.4) is the capability of the lumped reactions to direct all the carbon from R5P to dTTP. When we analyse the 4 lumped reactions with highest yield, we see that on the product side, the only compound other than inorganics and cofactor pairs like quinone/quinol, ATP/ADP-PI is dTTP. For the lumps with second highest yield, along with the pyruvate (C3) in the reactant side, we see AcCoA (C2) and CO_2 (C1) on the product side. In the 3rd highest yield case, fumarate (C4) replaces pyruvate on the reactant side, and succinate (C4) appears on the product side as a by-product. The lowest yield producing lumped reaction has fumarate and pyruvate on the reactant side, and has AcCoA, CO_2 and succinate on the product side of the equation. This signifies that the lump reactions with lower yield are losing carbon through those *core* metabolites and the number of carbons in these metabolites defines the yield. Having a lower yield does not necessarily mean that these lumped reactions are not useful for metabolic modelling, since they can be used under sub-optimal growth conditions or under conditions when growth is not the main physiological optimality criterion or to provide flexibility under mutation. The use of these alternatives and their physiological interpretation will be the subject of future studies.

4.3.3 Generating a metabolic model with lumpGEM

By generating subnetworks for GAM and diphosphate, lumpGEM managed to take into account all the components of biomass formulation both on the product and reactant sides. By testing for yield, we have shown that all the generated lumped reactions are capable of producing their target BBB_j. However, to produce biomass, these lumped reactions must be able carry flux under the same quasi steady state condition, (in modelling terms) in the same model with biomass as cellular objective. This requires generating a metabolic network composed of the defined *core* network, specifically Varma stoichiometry, lumped reactions, the transport and sink reactions defined in GEM. lumpGEM generated 673 unique lumped reactions for all 87 $BBBs$ and 2 reactions for diphosphate secretion by biomass. In order to test the lumpGEM results for optimum growth, we have selected the highest yield producing lumped reactions per BBB_j (if there are many, we chose 1 among them randomly). This network includes 56 cytosolic enzymatic (Varma network), 190 transport reactions, 88 lumpGEM output reactions along with 64 sinks (398 in total) and 158 unique metabolites. The metabolic network is capable of producing 0.944/hr specific growth rate with 10 mmol/gDWhr glucose uptake rate. This signifies that all the lumped reaction were capable of producing corresponding BBB_j successfully under the given condition simultaneously. The specific growth rate of GEM for the same condition is 0.997/hr, which is close to the biomass accumulation of the model generated with the output of lumpGEM. This shows that lumpGEM can be used to generate networks, which are small, but comprehensive and can mimic the GEM behaviour. It also proves that lumpGEM can generate lumped reactions from any part of the metabolism either handpicked (*ad hoc*) or systematically selected for metabolic modelling purposes.

4.3.4 Analysis on Compartmentalized Models, test case on *S. cerevisiae*

lumpGEM can be applied to any GEM with a proper biomass formulation, and in order to test its applicability, we used it on a eukaryotic, compartmentalized GEM, iMM904 [117] of *S. cerevisiae*. This yeast is one of the mostly studied unicellular organisms along with *E. coli*, and has many biotechnological applications [130]; thus making it a strong

candidate for modelling approaches. Applying lumpGEM on this yeast revealed the contribution of different possible precursors and cofactors for each of the biomass building blocks defined in GEM, moreover it revealed alternative subnetworks/lumped reactions for the same biomass building block. This can be interpreted as building 'Neidhardt style' tables for *S. cerevisiae.*

In order to define the core for *S. cerevisiae,* we used a labelling study from Christen et al. [131] and mapped the reactions of this MFA model with the iMM904 reactions. In addition, we mapped reactions of iMM904 that are in Varma network but not in C13 model and included them as *core* reactions for *S. cerevisiae* network. The generated core network is composed of 87 reactions and 67 unique metabolites along 2 compartments, cytoplasm and mitochondria. Following these steps, we have applied the lumpGEM algorithm to the GEM as described in Material Methods section. The main difference between *E. coli* and yeast GEMs are that yeast GEM is compartmentalized, however this does not bring any more complexity for lumpGEM since it treats the transport reactions between compartments as it treats uni-compartmentalized reactions. These transport reactions can be a part of the generated subnetworks for biomass building blocks and can participate in lumped reactions. The synthesis pathways of common biomass building blocks of yeast and *E. coli* such as amino acids are very similar. The sizes of the networks differ mainly from the transport reactions between the compartments. Another reason for the divergence is the non-core metabolites along the linear routes to biomass building blocks. There are different enzymes in these two organisms that are balancing these non-core metabolites. Different subnetworks do not necessarily produces different lumped reactions. When the synthesis pathway is in one compartment (mainly cytosol), the lumped reactions of *E. coli* and yeast are very similar. L-phenylalanine, L-methionine, L-serine L-cysteine and L-tyrosine are some examples of these similarities. Small differences for these overall reactions emerge from different cofactor usage. A similar behaviour is also observed for subnetworks including more than 1 compartment. The main difference between *E. coli* and yeast overall reactions emerges from metabolites in different compartments, which are also different due to the energetics cost of transport reactions.

Table 4-5: Lumped Reactions and Statistics for Amino Acids, Comparison between *E. coli* and yeast. The numbers (i:j:k) represent i, size of subnetworks for amino acids (2:1:1), j, the number of alternative subnetworks (2:1:1) and k, the number of unique lumped reactions (2:1:1), respectively. The first (i:j:k) belongs to *S. cerevisiae*, the second (i:j:k) belongs to *E. coli* for comparison. In this table the size corresponds only to smallest size subnetworks (S_{min}^{j}) for each BBB. Highlighted metabolites show the differences between alternative lumped reactions.

BIOMASS BUILDING BLOCK	LUMPED REACTIONS
L-ALANINE	2:1:1/2:1:1 H + NADPH + NH4 + PYR -> ALA-L + H2O + NADP
L-ARGININE	19:2:2/13:2:2 AKG_M + **5 ATP** + ATP_M + CO2 + 6 CO2_M + FADH2_M + **6 H2O** + 3 NADPH + NADPH_M + 4 NH4 + 3 OAA -> **5 ADP** + ADP_M + ARG-L + FAD_M + **7 H** + 2 H_M + 6 HCO3 + 3 NADP + NADP_M + 2 OAA_M + **5 PI** + PI_M + SUCC AKG_M + **8 ATP** + ATP_M + CO2 + 6 CO2_M + FADH2_M + **9 H2O** + 3 NADH + NADPH_M + 4 NH4 + 3 OAA -> **8 ADP** + ADP_M + ARG-L + FAD_M + **10 H** + 2 H_M + 6 HCO3 + 3 NAD + NADP_M + 2 OAA_M + **8 PI** + PI_M + SUCC
L-ASPARAGINE	7:2:2/5:2:2 **3 ATP** + CO2 + **2 H2O** + NADPH + 2 NH4 + OAA -> **3 ADP** + ASN-L + **3 H** + HCO3 + NADP + **3 PI** **4 ATP** + CO2 + **3 H2O** + NADH + 2 NH4 + OAA -> **4 ADP** + ASN-L + **4 H** + HCO3 + NAD + **4 PI**
L-ASPARTATE	3:1:1/2:1:1 CO2 + NADPH + NH4 + OAA -> ASP-L + HCO3 + NADP
L-CYSTEINE	14:2:2/16:6:2 3PG + ACCOA + AMP + **3 ATP** + NAD + 5 NADPH + NH4 + PPI + SO4 -> AC + **4 ADP** + COA + CYS-L + **H2O** + NADH + 5 NADP + **5 PI** 3PG + ACCOA + AMP + **4 ATP** + NAD + 5 NADPH + NH4 + PPI + SO4 -> AC + **5 ADP** + COA + CYS-L + **H** + NADH + 5 NADP + **6 PI**
L-GLUTAMINE	7:1:1/2:2:2 AKG + ATP + CIT_M + 2 CO2 + 2 CO2_M + 2 FADH2_M + 2 FUM + 4 H2O + MAL-L + NADPH + 2 NH4 -> ADP + CIT + 2 FAD_M + GLN-L + H2O_M + 4 H + 4 HCO3 + MAL-L_M + NADP + PI + 2 SUCC
L-GLUTAMATE	7:1:1/1:1:1 AKG + H + NADPH + NH4 -> GLU-L + H2O + NADP

GLYCINE	9:1:1/8:2:2 CIT_M + 2 CO2 + 2 CO2_M + 2 FADH2_M + 2 FUM + GLX + 4 H2O + MAL-L + NADPH + NH4 -> CIT + 2 FAD_M + GLY + H2O_M + 3 H + 4 HCO3 + MAL-L_M + NADP + 2 SUCC
L-HISTIDINE	25:9:3/21:12:3 3PG + **6 ATP** + **3 H2O** + MLTHF + **3 NAD** + **2 NADPH** + 4 NH4 + OAA + R5P -> **6 ADP** + FUM + 12 H + HIS-L + **3 NADH** + **2 NADP** + 8 PI + SER-L + THF 3PG + **6 ATP** + **3 H2O** + MLTHF + **4 NAD** + **3 NADPH** + 4 NH4 + OAA + R5P -> **6 ADP** + FUM + 12 H + HIS-L + **4 NADH** + **3 NADP** + 8 PI + SER-L + THF 3PG + 9 ATP + **6 H2O** + MLTHF + **NAD** + 4 NH4 + OAA + R5P -> **9 ADP** + FUM + 15 H + HIS-L + **NADH** + **11 PI** + SER-L + THF
L-ISOLEUCINE	17:2:2/12:1:1 AC_M + 2 ATP + 2 CO2_M + H2O + H + 2 H_M + **4 NADPH** + NADPH_M + NH4 + OAA + PYR_M -> AC + 2 ADP + CO2 + H2O_M + 2 HCO3 + ILE-L + **4 NADP** + NADP_M + 2 PI AC_M + 2 ATP + 2 CO2_M + H2O + H + 2 H_M **+ NADH + 3 NADPH** + NADPH_M + NH4 + OAA + PYR_M -> AC + 2 ADP + CO2 + H2O_M + 2 HCO3 + ILE-L + **NAD + 3 NADP** + NADP_M + 2 PI
L-LEUCINE	13:1:1/10:1:1 ACCOA_M + 3 CO2_M + H2O + H_M + NAD + NADPH + NADPH_M + NH4 + 2 PYR_M -> 3 CO2 + COA_M + H + 2 HCO3 + LEU-L + NADH + NADP + NADP_M
L-LYSINE	17:2:2/11:1:1 ACCOA_M + AKG_M + 2 ATP + CO2 + 4 CO2_M + 4 FADH2_M + 4 FUM + 7 H2O + **NAD + NAD_M + 4 NADPH** + 2 NH4 -> 2 ADP + COA_M + 4 FAD_M + H2O_M + 4 H + H_M + 6 HCO3 + LYS-L + **NADH + NADH_M + 4 NADP** + 2 PI + 4 SUCC ACCOA_M + AKG_M + 2 ATP + CO2 + 4 CO2_M + 4 FADH2_M + 4 FUM + 7 H2O + **NAD_M + 3 NADPH** + 2 NH4 -> 2 ADP + COA_M + 4 FAD_M + H2O_M + 4 H + H_M + 6 HCO3 + LYS-L + **NADH_M + 3 NADP** + 2 PI + 4 SUCC
L-METHIONINE	21:4:4/25:6:2 3PG + ACCOA + AMP + **4 ATP** + CO2 + 3 H + MLTHF + **8 NADPH** + 2 NH4 + OAA + PPI + SO4 -> AC + **5 ADP** + COA + H2O + HCO3 + MET-L + **8 NADP** + 6 PI + SER-L + THF 3PG + ACCOA + AMP + **4 ATP** + CO2 + 3 H + MLTHF + **NAD + 9 NADPH** + 2 NH4 + OAA + PPI + SO4 ->

	AC + **5 ADP** + COA + H2O + HCO3 + MET-L + **NADH + 9 NADP** + 6 PI + SER-L + THF 3PG + ACCOA + AMP + **5 ATP** + CO2 + 2 H + MLTHF + **8 NADPH** + 2 NH4 + OAA + PPI + SO4 -> AC + **6 ADP** + COA + HCO3 + MET-L + **8 NADP** + **7 PI** + SER-L + THF 3PG + ACCOA + AMP + **5 ATP** + CO2 + 2 H + MLTHF + **NAD** + **9 NADPH** + 2 NH4 + OAA + PPI + SO4 -> AC + **6 ADP** + COA + HCO3 + MET-L + **NADH** + **9 NADP** + 7 PI + SER-L + THF
L-PHENYLALANINE	11:1:1/11:1:1 ATP + E4P + 2 NADPH + NH4 + 2 PEP -> ADP + CO2 + 2 H2O + H + 2 NADP + PHE-L + 4 PI
L-PROLINE	10:2:2/5:1:1 AKG + ATP + CIT_M + 2 CO2 + 2 CO2_M + 2 FADH2_M + 2 FUM + 3 H2O + MAL-L + **3 NADPH** + NH4 -> ADP + CIT + 2 FAD_M + H2O_M + 2 H + 4 HCO3 + MAL-L_M + **3 NADP** + PI + PRO-L + 2 SUCC AKG + ATP + CIT_M + 2 CO2 + 2 CO2_M + 2 FADH2_M + 2 FUM + 3 H2O + MAL-L + **NADH + 2 NADPH** + NH4 -> ADP + CIT + 2 FAD_M + H2O_M + 2 H + 4 HCO3 + MAL-L_M + **NAD + 2 NADP** + PI + PRO-L + 2 SUCC
L-SERINE	4:1:1/4:1:1 3PG + NAD + NADPH + NH4 -> H + NADH + NADP + PI + SER-L
L-THREONINE	8:2:2/7:1:1 2 ATP + CO2 + H2O + H + **3 NADPH** + NH4 + OAA -> 2 ADP + HCO3 + **3 NADP** + 2 PI + THR-L 2 ATP + CO2 + H2O + H + **NADH + 2 NADPH** + NH4 + OAA -> 2 ADP + HCO3 + **NAD** + 2 NADP + 2 PI + THR-L
L-TRYPTOPHAN	19:2:2/17:2:2 3PG + **4 ATP** + E4P + **NAD + 2 NADPH** + 2 NH4 + 2 PEP + R5P -> 3IG3P + **4 ADP** + CO2 + **7 H** + **NADH + 2 NADP** + **8 PI** + PYR + SER-L 3PG + **5 ATP** + E4P + H2O + **NADPH** + 2 NH4 + 2 PEP + R5P -> 3IG3P + **5 ADP** + CO2 + **8 H** + **NADP** + **9 PI** + PYR + SER-L
L-TYROSINE	11:2:2/11:1:1 ATP + E4P + **NAD** + **2 NADPH** + NH4 + 2 PEP -> ADP + CO2 + H2O + 2 H + **NADH** + **2 NADP** + 4 PI + TYR-L ATP + E4P + **NADPH** + NH4 + 2 PEP -> ADP + CO2 + H2O + 2 H + **NADP** + 4 PI + TYR-L
L-VALINE	9:1:1/5:1:1 AC_M + 3 CO2_M + H2O + 2 H_M + NADPH + NADPH_M + NH4 + 2 PYR_M -> AC + 2 CO2 + H2O_M + H + 2 HCO3 + NADP + NADP_M + VAL-L

4.4 Conclusion

In this chapter, we mainly focused on the biosynthetic pathways of *E. coli* and yeast, and by using an *ad hoc* built central carbon network, we generated possible production routes for all biomass building blocks defined in GEMs. This approach allowed us to re-define the pathways, and we discovered that the contribution of many enzymes that are labelled with different subsystems contributing for the biosynthesis of many BBBs. Moreover, by lumping the generated subnetworks made it possible to study the individual contribution of core metabolites and cofactors for synthesis of each BBB. The lumping method is a very promising method for many analysis, experimental studies like MFA or *in silico* studies such as FBA, TFA or Atom Mapping since it deduces the precursor metabolites from the central carbon network.

lumpGEM can also be used to build synthesis pathways for any metabolite defined in the metabolic network. This makes it a strong tool to study the characteristics of industrial chemical production strains, since it identifies all the enzymes, either linearly connected or nested that participate in the biosynthesis of the target compound.

Apart from pathway and subsystems/subnetworks analysis, lumping strategy reduces the complexity of the networks significantly. This property of lumpGEM became the key factor for us to build representative reduced models from GEMs, which is the topic of the next chapter.

Chapter 5 - *redGEM*: Systematic reduction of genome-scale metabolic reconstructions for development of core metabolic models

5.1 Introduction

In the previous chapters, we discussed about the stoichiometric models and their evolution to Genome Scale Metabolic Reconstructions (GEM), which encapsulate all known biochemistry that takes place in the organisms by gene to protein to reaction (GPRs) associations [112]. Since the first Genome Scale model developed in 2000 [115], the number of annotated genomes and the corresponding genome scale metabolic reconstruction increased tremendously [24,132,133]. This enabled the researches to study the metabolic networks of different organisms, from archaea [134] to human.

With increasing popularity of Genome Scale Metabolic Reconstructions (GEMs), different techniques to analyse these networks have been proposed [135,136]. Flux Balance Analysis (FBA) [33] and Thermodynamic-based Flux Analysis (TFA) [12,13,15,137] for the integration of available information with GEMs are invaluable tools to study metabolic networks as discussed in Chapter 1. However, both FBA, and TFA cannot capture the dynamic response of metabolic networks, which requires integration of detailed enzyme kinetics and regulations [16]. Hatzimanikatis and colleagues have recently developed a framework that utilizes FBA, TFA and generates kinetic models without sacrificing stoichiometric, thermodynamic and physiological constraints [73]. They have employed methods that are also applicable to GEMs, however due to complexity of the

GEMs and to focus on certain parts of the metabolism, they used a reduced model for the analysis.

As the quality and the size of the models increase with better annotation, the complexity of the mathematical representations of the models also increase. Keng et al. [34] observed that majority of the studies and applications using metabolic models have still revolved around the central metabolism and around "reduced" models instead of genome-scale models, indicating that the full potential of GEMs remains largely untapped [138-140]. Moreover, these models have often been reduced *ad hoc,* with different criteria and aims, which have not been consistently and explicitly justified [109,141,142]. Therefore, an approach to focus on chosen metabolic subsystems and yet retain the linkages and knowledge captured in genome-scale reconstructions would help us to manage the complexity and to yield better insights in connecting the metabolic model to actual cellular physiology.

In this study, we have developed redGEM, a systematic model reduction framework for constructing core metabolic models from GEMs. In redGEM, we use as inputs: (i) metabolic subsystems that are of interest for a physiology under study; (ii) information about utilized carbon source; and (iii) available physiological data. After a series of computational processes, we generate a reduced model that is consistent with the original GEM in terms of flux profiles, essential genes and reactions. We applied redGEM on the latest GEM of *E. coli iJO1366* under both aerobic and anaerobic conditions with glucose as the sole carbon source and generated *E. coli* ***r****iJO1366*.

5.2 Material Methods

We applied redGEM algorithm on the latest genome scale model of *E. coli* iJO1366 [6], which is composed of 2253 enzymatic reactions (along with transporters), 335 boundary reactions and 1139 unique metabolites across cytoplasm, periplasm and extracellular media. We constrained the model for aerobic conditions with glucose as the sole carbon source.

5.2.1 Preliminary Definitions

S_i: Core subsystem *i* that is selected/defined by the user.

M^{S_i}: Metabolites that belong to subsystems S_i.

R^{S_i}: Reactions that belong to subsystems S_i.

Degree of linkage D: The length of paths between two subsystems. It corresponds to the number of reactions that link subsystems $S_i and\ S_j$.

R_{ij}^{D}: The reactions in all paths of length D between the subsystems $S_i$and S_j; these reactions do not belong to either R^{S_i}or R^{S_j}.

M_{ij}^{D}: The metabolites that are intermediates in all paths of length D between the subsystems $S_i$and S_j; these metabolites do not belong to either M^{S_i} or M^{S_j}.

Postulate 1: Reactions that belong to R_{ij}^{D} and metabolites that belong to M_{ij}^{D} can belong to any of the subsystems S_m with $m \neq i\ and\ m \neq j$.

Postulate 2: Some of the reactions in R_{ij}^{D+n} can belong to R_{ij}^{D}. Reactions in $R_{ij}^{D=n}$ that do not belong in any other $R_{ij}^{D=1,2,\cdots}$ (where $D \neq n$) are called *unique reactions* for the degree of linkage *D*.

Postulate 3: R_{ij}^{D} and M_{ij}^{D} captures the linking between the non-common metabolites of $S_i$and S_j, however it cannot capture the intra-connections between the metabolites of the same subsystem or the metabolites that are shared between $S_i$and S_j.

R_{ii}^{D}: The reactions in all paths of length D that intra-connects the metabolites of the subsystem S_i.

M_{ii}^{D}: The intermediate metabolites in all paths of length D that intra-connects the metabolites of the subsystem S_i.

R^T: Reactions where only M^{S_i}, M_{ij}^D and M_{ii}^D, participate and do not belong to R^{S_i}, R_{ij}^D and R_{ii}^D.

Postulate 4: R^T is composed of reactions that only cofactor pairs, small metabolites and inorganics participate. All the other reactions that include other core metabolites (along with cofactor pairs, small metabolites and inorganics) will be a part of R^{S_i}, R_{ij}^D or R_{ii}^D.

Core Network, CN^D: The core network for redGEM that is composed of metabolites M^{S_i}, M_{ij}^D and M_{ii}^D, and of reactions R^{S_i}, R_{ij}^D, R_{ii}^D and R^T.

We can generate the core network from the chosen subsystems using the minimum distance between the chosen subsystems and reports the linking reactions and metabolites. In this case, the degree of linking *D* is the minimum distance between $S_i$and S_j.

$L_{min,ij}$*:* The length of the shortest path between the subsystems $S_i$and S_j.

$R_{i,j}^{L_{min+n,ij}}$: The reactions that connect the subsystems $S_i$and S_j with a path of length $L_{min+n,ij}$ in where *n* is a user defined parameter.

$M_{ij}^{L_{min+n,ij}}$: The metabolites that does not belong to either $S_i$or $S_j$and are intermediates of the path of length $L_{min+n,ij}$ in between these two subsystems.

Postulate 5: If $L_{min,ij} = 1$ then $R, M_{ij}^{L_{min,ij}}$ becomes R, M_{ij}^1, this also results in R, M_{ij}^D = $R, M_{ij}^{L_{min+(D-1),ij}}$.

5.2.2 redGEM Parameters

redGEM uses the following inputs and parameters:

1. The starting subsystems or sets of reactions/metabolites defined by the user.
2. Media conditions (aerobic/anaerobic, nitrogen limited, etc.).
3. Possible carbon sources for the studied physiology.
4. Possible by-products or relevant extracellular metabolites. Together with possible carbon sources, these metabolites form a new subsystem that redGEM names

as Extracellular Subsystem, this subsystem is treated as other subsystems defined in Step 1 above.

5. Organism specific cofactor pairs.
6. Degree of linking *D* defined by the user.

5.2.3 redGEM Workflow

1.a Choose subsystems (or list of reactions and metabolites, such as synthesis pathway of a target molecule) based on the studied physiology and the part of the metabolism under interest.

1.b Derive a new stoichiometric matrix that excludes all cofactor pairs, small metabolites and inorganics.

1.c Identify R^S, R_{ij}^D, R_{ii}^D, R^T, and M^S, M_{ij}^D and M_{ii}^D for all subsystem pairs except Extracellular Subsystem.

- Perform a graph search on the new stoichiometric matrix.
- This will find all the links up to degree D between each subsystem pairs $S_i and\ S_j$, and will not find any reaction or metabolites between two subsystems if $L_{min,ij} > D$.

1.d To connect all Extracellular Subsystem metabolites to other subsystems, find all reactions $R_{i,j}^{L_{min+n,ij}}$ and all metabolites $M_{ij}^{L_{min+n,ij}}$, with *n* as defined by the user.

- If the length of shortest path between a metabolite and S_i is bigger than 1, then:
- $number\ of\ R_{ij}^{L_{min+n,ij}} \geq\ number\ of\ R_{i,j}^{n+1}$
- $number\ of\ of\ M_{ij}^{L_{min+n,ij}} \geq\ number\ of\ M_{ij}^{n+1}$

The core carbon network is defined as all the reactions and metabolites in M^S, M_{ij}^D and M_{ii}^D (all *i, j* pairs), R^{S^i}, R_{ij}^D, R_{ii}^D (all *i, j* pairs) and R^T.

5.2.4 Formulation of lumped biosynthetic reactions for biomass building blocks

We used the formulation as described Chapter 4 and generated pathways for all biomass building blocks (BBB) as they are defined in GEM by lumpGEM. We have generated subnetworks up to minimum subnetwork size plus 3 for each BBB. Then we have calculated the unique lumped reactions for all the BBBs, and used these lumped reactions for further validation and other analysis.

5.2.5 Validation

We validate the consistency between GEM and rGEM performing the following consistency checks:

i. Theoretical maximum biomass and other by-product of interest yield of redGEM and GEM growing on same carbon source.
    a. Under aerobic and anaerobic conditions for the organisms that can grow under both conditions.

ii. Essentiality of the common genes.
    a. Perform single gene deletions in order to see consistency between redGEM and GEM.
        i. Perform gene essentiality with FBA and with TFA.

iii. Allowable flux ranges of the common reactions between rGEM and GEM.
    a. By performing Flux Variability Analysis (FVA) on both rGEM and GEM, and compare the common reactions with FBA and TFA.
        i. Perform FVA with (FBA) and without thermodynamic constraints (TVA)

iv. Allowable concentration ranges for all metabolites with thermodynamics information.
    a. TFA formulation allows the integration of steady-state concentrations of metabolites since it constrains the model through Gibbs free energy of reaction.

b. Performed i) CVA (Concentration Variability Analysis) for GEM and redGEM and compare the allowable ranges for common metabolites between the two models and ii) TVA to compare the ΔG_{rxn} range of reactions.

While these are the basic consistency tests, one could define additional checks, which can be specific to the organism and problem under study. We recommend that in all cases one should perform the checks using FBA and TFA, i.e. with and without thermodynamics constraints.

5.3 Results and Discussion

We performed a reduction of the *E. coli* GEM iJO1366 using redGEM. We selected 6 central carbon metabolism subsystems (glycolysis, pentose phosphate pathway, citric acid cycle, glyoxylate cycle, pyruvate metabolism, and oxidative phosphorylation) as they are defined in original *E. coli* GEM. In addition, we have included all the reactions that use quinone/quinol pool metabolites (Ubiquinone/ubiquinol, menaquinone/menaquinol, 2-dimethyl menaquinone/2- dimethyl menaquinol for *E. coli*) in oxidative phosphorylation subsystem to capture the coupling between the core carbon metabolism and energy/redox metabolism. Some of those reactions had different subsystem definition in original GEM. These subsystems include a total of 184 reactions and 125 metabolites (Appendix Table A.3). We next redefined the content of the subsystems by performing a tightening analysis to identify the R^T (See Materials Methods) reactions unique to each subsystem. This analysis established that there are many reactions in GEM that include metabolites that belong to only a specific subsystem but are labeled to a different subsystem in the original GEM definition (Table 5.1). Some of the reactions defined in R^T are common between subsystems, since the subsystems share many metabolites, especially cofactor pairs such as ATP/ADP, NAD^+/NADH etc.

Table 5-1: Statistics on Starting Subsystems with tightening reactions, R^T

Subsystems	Metabolites	Reactions	Tightening Reactions
Citric Acid Cycle	24	10	6
Pentose Phosphate Pathway	21	12	2
Glycolysis/Gluconeogenesis	35	22	17

Pyruvate Metabolism	22	10	3
Glyoxylate Metabolism	13	4	3
Oxidative Phosphorylation	72	70	24
Media Composition	10	10	-

Unlike other reduction approaches such as *Network Reducer* [143], which reduces the GEM by removing reactions through Flux Variability Analysis (FVA), redGEM performs the reduction of GEMs by expansion of core network around the selected subsystems. The network expansion by directed graph search finds metabolites and reactions between subsystems in a pairwise manner. It performs pairwise connections between the metabolites of the same subsystems. By definition, the linking procedure is performed through metabolites. Hence, R^{T,S^i} does not change the calculated pairwise connections between subsystems since it does not add any new metabolite to the subsystems. redGEM calculates M^S, M_{ij}^D and M_{ii}^D (all pairs *i, j*), R^{S_i}, R_{ij}^D, R_{ii}^D (all pairs *i, j*), which overall define the core network with respect to selected degree of connection parameter *D.* As a final step, redGEM performs once more tightening, and scans through every reaction in GEM to identify the reactions R^T, which are not captured by R^{S_i}, R_{ij}^D, R_{ii}^D (all pairs *i, j*) but include only M^S, M_{ij}^D and M_{ii}^D. This procedure finalizes the steps that define the final core network for further analysis for redGEM/lumpGEM.

Table 5-2: The statistics of different Core Networks. The reported values for metabolites are compartmentalized, i.e. pyruvate cytoplasmic and cytosol periplasmic are reported as different metabolites.

Degree of Connection	**# of Metabolites**	**# of Reactions**
D=1	156	243
D=2	202	289
D=3	224	320
D=4	268	358
D=5	391	544
D=6	496	689

For D=1, redGEM captures many reactions that are part of many *ad hoc* built models, such as malic enzymes 1-2 between glycolysis and TCA cycle that connect L-malate to pyruvate, phosphoenolpyruvate carboxylase and phosphoenolpyruvate carboxykinase

that connect oxaloacetate and phosphoenolpyruvate. Moreover, it captures many other reactions, such as 2 types of L-aspartate oxidases, which are using quinone/quinol co-factor pairs and labeled as electron transport chains reactions. There are 2 more L-aspartate oxidase reactions that are added to the D=1 core network by redGEM. One is using O_2/H_2O_2 and the other one is using fumarate/succinate as cofactor pairs. These reactions are captured by R_{ii}^{D} and R^{T} simultaneously. Finally redGEM added 10 more reactions that only have the cofactors belonging to D=1 core network in their stoichiometry, such as NAD^+ kinase, NADP phosphatase, adenylate kinase, nucleoside-triphosphatase etc. as a part of R^{T}.

By starting from 7 subsystems (including extracellular metabolites as exchange), the network expansion procedure results in capturing reactions as core from 33 different subsystems for D=6 (Table 4.3). In GEM, there are 36 subsystems, which signifies that only 6 steps expansion captures reactions from ~90% of all subsystems defined in GEM, thus showing the tight connections between metabolites/subsystems in the network.

Table 5-3: The subsystems that can be reached from starting subsystems in 6 steps

SUBSYSTEMS REPRESENTED IN D=6 CORE NETWORK	
Alanine and Aspartate Metabolism	Lipopolysaccharide Biosynthesis / Recycling
Alternate Carbon Metabolism	Methionine Metabolism
Anaplerotic Reactions	Methylglyoxal Metabolism
Arginine and Proline Metabolism	Murein Recycling
Cell Envelope Biosynthesis	Nitrogen Metabolism
Citric Acid Cycle	Nucleotide Salvage Pathway
Cofactor and Prosthetic Group Biosynthesis	Oxidative Phosphorylation
Cysteine Metabolism	Pentose Phosphate Pathway
Folate Metabolism	Purine and Pyrimidine Biosynthesis
Glutamate Metabolism	Pyruvate Metabolism
Glycerophospholipid Metabolism	Threonine and Lysine Metabolism
Glycine and Serine Metabolism	Transport, Inner Membrane
Glycolysis/Gluconeogenesis	Transport, Outer Membrane Porin
Glyoxylate Metabolism	Tyrosine, Tryptophan, and Phenylalanine Metabolism
Histidine Metabolism	Unassigned – No Subsystem
Inorganic Ion Transport and Metabolism	Valine, Leucine, and Isoleucine Metabolism

5.3.1 Generation of Lumped Reactions for Biomass Building Blocks from Core Carbon Network

The wild-type biomass reaction of the iJO1366 model contains 102 biomass building blocks. The size and the complexity of the composition makes it necessary to develop techniques to keep the information stored in the biomass formulation, but yet reduce the size of the network significantly. Methods, such as graph-search algorithms can be used for identification of biosynthetic routes between two metabolites in metabolic networks [144]. However, these graph theory based approaches cannot be used for our purposes due to the following limitations/issues: *i) they do not make use nor obey mass conservation; hence the pathways they generate are not guaranteed to be able to carry flux in metabolic network or to be elementally balanced. ii) They cannot manage non-linear branched pathways.* As explained in Chapter 4, we have developed lumpGEM to overcome these limitations. lumpGEM identifies subnetworks for each biomass building block using the *core* metabolites as precursors and it generates lumped reactions for each subnetwork. Moreover, it obeys to the mass conservation constraints, thus preventing the generation of lumped reactions, which cannot carry flux in the metabolic networks. In Chapter 4, the core metabolites were the ones belonging to the Varma network that was built with an *ad hoc* manner. redGEM provides the new core metabolites by defining core networks based on degree of connection parameter, D. For D=1, by maximizing the number of non-active *non-core* reactions In GEM, lumpGEM generated a 17 reactions subnetwork to synthesize histidine from *core* carbon metabolites (Figure 5.2). In Chapter 4, lumpGEM generated a 21 reactions subnetwork for histidine synthesis. This shows that redGEM defined at least 4 reactions from 21 reactions as core. Like in the example in Chapter 4, histidine is synthesized from ribose-5-phosphate, a precursor from pentose phosphate pathway. The linear pathway from this *core* metabolite to histidine is composed of 10 steps. However, due to the mass balance constraint, two metabolites, 1-(5-Phosphoribosyl)-5-amino-4-imidazolecarboxamide and L-Glutamine cannot be balanced in a network that is composed of *core* reactions and the linear pathway from ribose-5-phophate to histidine. These metabolites are balanced in the network by other *non-core* reactions. Hence, the generated sets of reactions are not lin-

ear routes from precursor metabolites to biomass building blocks, but branched, *balanced subnetworks.*

Using lumpGEM, we replicated all the biosynthetic pathways in databases such as EcoCyc [23], either as a part of subnetworks or the pathway itself, in addition we also identify subnetworks that can qualify as alternative biosynthetic pathways. *E. coli* is well-known to be robust against deletions by having many duplicate genes and alternate pathways[145]. Some of these routes may not be active due to energetics or regulatory constraints but using our method here can help us to map out these possible alternate pathways completely and also derive different biosynthetic lumped reactions. The introduction of such lumped biosynthetic reactions simplifies the core models considerably and allows the use of these models in important computational analysis methods such as dynamic FBA [146] extreme pathway analysis [147,148] and elementary flux modes [129,149], as well as TFA formulations and kinetic modeling. lumpGEM generated 1093 subnetworks and 246 unique lumped reactions for 79 biomass building blocks. The GEM has 102 BBBs, thus indicating that 15 BBBs can be produced within the D=1 core network.

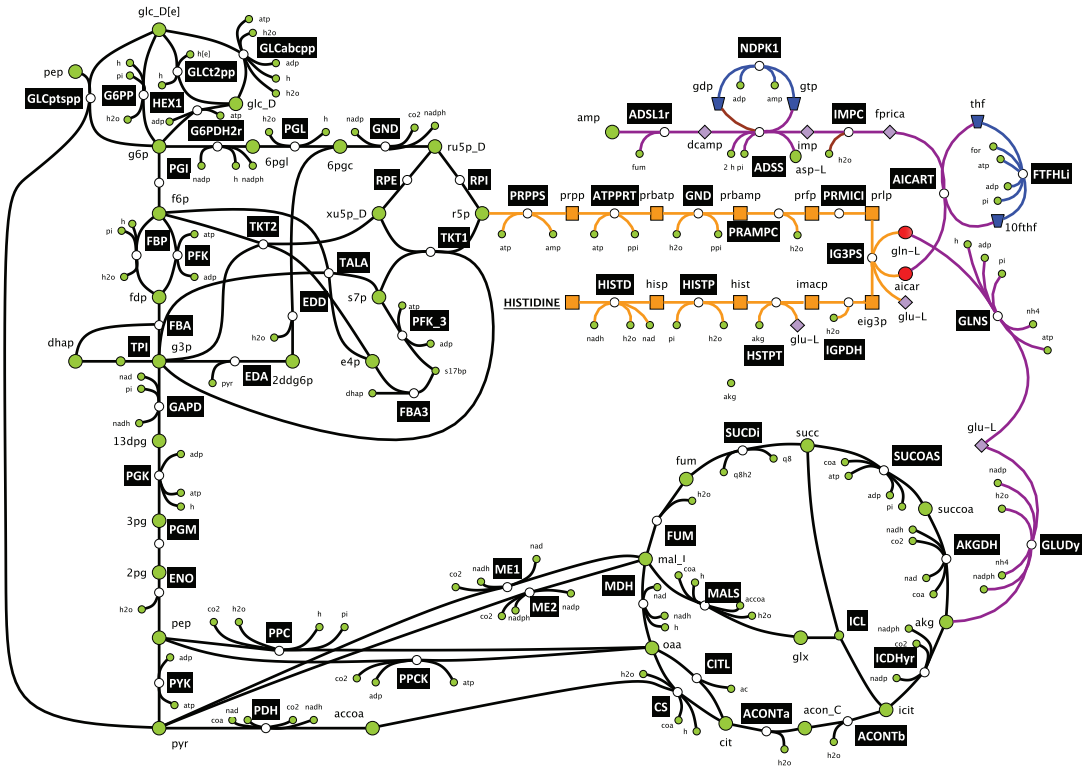

Figure 5-1: The synthesis of histidine from core carbon network. Histidine synthesis starts from ribose-5-phosphate (R5P) from Pentose Phosphate Pathway, and consists of 10 steps. The *non-core* metabolites along the pathway are balanced by the reactions other than the linear pathway for histidine synthesis. Metabolites colored with green are *core*; with red are *non-core* metabolites that are balanced by reactions, which are non-core, and not a part of the linear route from R5P to histidine (Not all reactions of core network are shown). Orange reactions form the linear pathways for histidine as defined in databases. Purple reactions are balancing the non-core metabolites and blue reactions are balancing the non-core metabolites appearing in purple reactions.

All those 246 lumped reactions are generated under aerobic conditions. For some biomass building blocks, it is possible that the all the alternatives for S_{min} subnetworks are using molecular oxygen, thus cannot carry flux under anaerobic conditions. This necessitates the generation of lumped reactions without any oxygen in the media. lumpGEM generated only 23 new lumped reactions for anaerobic case, for 7 metabolites, namely biotin, heme O, lipoate (protein bound), phosphatidylethanolamine (dioctadec-11-enoyl, n-C18:1), phosphatidylglycerol (dioctadec-11-enoyl, n-C18:1),

protoheme and cardiolipin (tetraoctadec-11-enoyl, n-C18:1). All the other lumped reactions generated for anaerobic case are a subset of the 246 lumped reactions for aerobic conditions.

5.3.2 Validation

5.3.2.1 Maximum biomass under different carbon sources

One of the most important criteria for the rGEM validation is the maximum biomass production. We performed biomass maximization with FBA and TFA. With all 269 (246+23) lumped reactions, maximum specific growth rate of the rGEM is the same as GEM's μ_{max}, 0.99 hr^{-1} with 10-mmol/gDWhr glucose uptake rate under aerobic conditions both with FBA and TFA. The anaerobic specific growth rate of GEM with the same carbon source for FBA is ~0.67/hr and with thermodynamic constraints (TFA) it drops to 0.27/hr. rGEM grows with 0.27/hr specific growth rate both with FBA and TFA. When we analyzed the discrepancy between the FBA and TFA growth rates for GEM, we saw that the difference is emerging from reactions that use molecular oxygen in GEM. These oxygen-using reactions do not belong to oxidative phosphorylation or ETC reactions, and are not a part of rGEM network. Moreover, the standard Gibbs free energy of those reactions range from 19kcal/mol to 294 kcal/mol in the oxygen producing direction[40] and thermodynamically infeasible, except for 5 reactions which are mainly degradation of hydrogen peroxide and superoxide anion. These 5 reactions have no effect on growth rate.

The reduction procedure can be performed on selected carbon source, and growth on a different carbon source may require different minimal subnetworks for the same biomass building blocks. Thus, to build a consistent rGEM, the lumping procedure should be repeated.

5.3.2.2 Gene Essentiality Comparison Between GEM and rGEM

One of the most common analyses for genome-scale models is *in silico* gene deletion (knockout) experiments to *i) identify essential and nonessential genes, ii) study the gene*

deletion impact on the organism physiology, iii) develop strategies for metabolic engineering [150]. Consistency of gene knockouts between rGEM and GEM is another important corroboration for the reliability of the reduction procedure. *iJO1366*, generated with D=1, shares 307 genes with GEM, and among these 307 genes, 26 of them are essential. Out of these 26 genes, 23 of them are also essential in GEM. 2 out of 3 genes do not have an effect on the maximum theoretical yield of *E. coli* under aerobic, minimal glucose medium in GEM. The first case is the gene transcribing thioredoxin reductase enzyme, which interconverts NADPH to NADP by using oxidized thioredoxin and reduced thioredoxin as cofactor pairs. This reaction is not essential in GEM, however it is essential in redGEM, since the cofactor pairs oxidized thioredoxin and reduced thioredoxin are participating in lumped reactions, and due to flux coupling, the reaction that thioredoxin reductase catalyzes becomes indispensible. Building alternative lumped reactions that do not use this cofactor pair may result in non-essentiality for this gene, however, lumped reactions built from S_{min} do not make this gene non-essential. The second discrepancy of the responses to gene deletion between GEM and rGEM is the gene transcribing Glutamate dehydrogenease, which shows a different behaviour compared to thioredoxin reductase enzyme. The reaction it catalyzes is the only reaction that synthesizes glutamate in the rGEM, and knocking out this enzyme automatically results in no specific growth. Knocking out this enzyme in GEM results in a growth drop about 3.3 percent, and alternative synthesis pathways for glutamate in GEM abolishes the essentiality. The deletion of adenylate kinase is the third discrepancy between GEM and rGEM. Knocking out this gene does not result in any drop in growth for GEM, however it prevents the cellular growth in rGEM. The reason for this discrepancy is similar to the case of Thioredoxin Reductase, i.e. the loss of alternative reactions/pathways that can complement this deletion. These reactions/pathways can be part of the lumped reactions, but this still reduces the flexibility of the network.

5.3.2.3 Flux Variability Analysis – Comparison between redGEM and GEM

To further validate the model, we compared the physiologically allowable flux ranges by performing Thermodynamics-based Variability Analysis (TVA) for the reactions that are

common between rGEM and GEM. Even though most of the common intracellular reactions have consistent flux ranges, there are some reactions in the core model with reduced flux variability as compared to the genome-scale model. The variability of the reactions in the subsystems Glycolysis/Gluconeogenesis, Pentose Phosphate Pathway, and Citric Acid Cycle are close to variability in GEM, due to the less alternative reactions for these reactions in GEM. Reactions that belong to Pyruvate Metabolism and Electron Transport Chains (ETC) show a higher variability in GEM compared to redGEM, due to the alternative reactions that use the metabolites that participate in those reactions. The main difference between GEM and rGEM emerges from directionalities, since rGEM is more constrained, some reactions, such as LDH-D (Lactate Dehydrogenase) become unidirectional. Moreover, as we discussed in the case of essentiality studies, the integration of reactions into lumped reactions reduces the flexibility of the flow in the network.

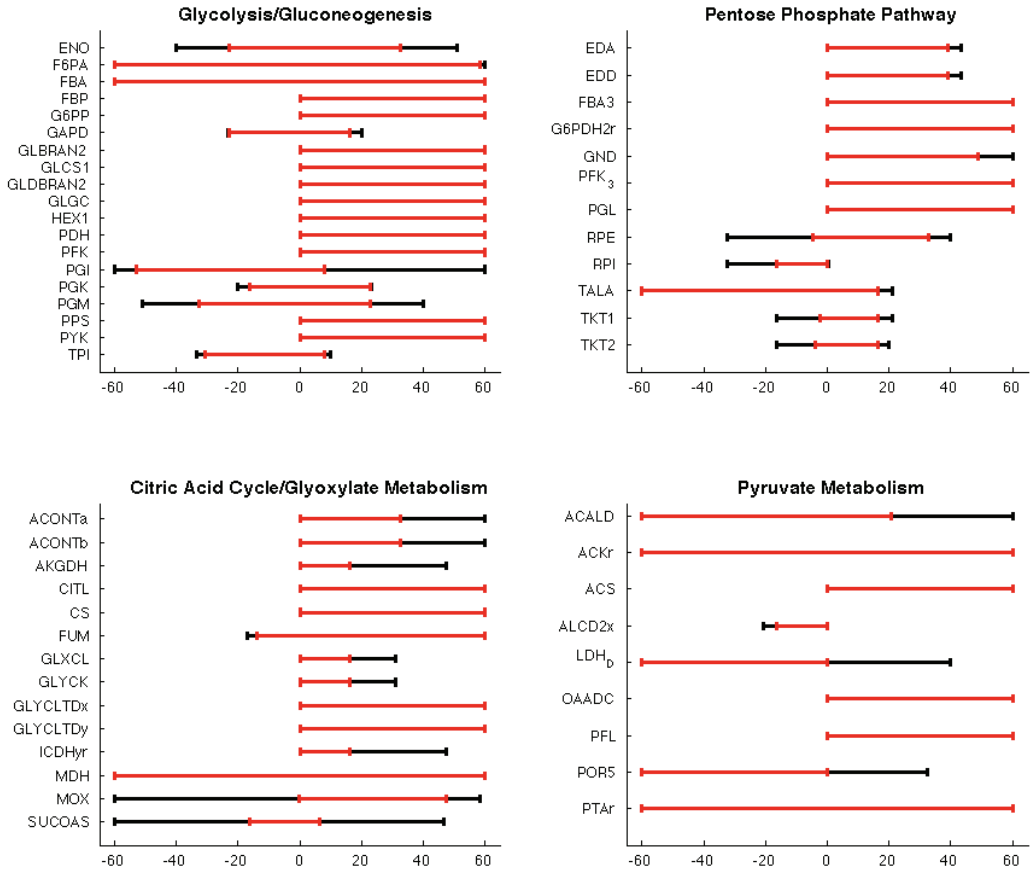

Figure 5-2: Flux variability of reactions in starting subsystems in D=1 model compared to corresponding reactions in GEM. The red lines represents FVA for redGEM, black lines represents TVA for GEM. There is not any reaction in rGEM that has a wider range than corresponding GEM reaction. Thus, for the reactions that do not have the black line have the same range for rGEM and GEM. Maximum flux bounds are between -60 to 60 mmol/gDWhr, since the uptake of glucose is fixed to 10 mmol/gDWhr, and the maximum allowable flux in the network cannot exceed 10 mmol/gDWhr times 6, which is the number of carbon in glucose. Variability of ETC reactions are reported in Appendix Table A.4.

Then, we performed a Concentration Variability Analysis (CVA) on common metabolites between GEM and rGEM (Figure 6.3). Almost all metabolites have the same allowable ranges, however there are a few exceptions. Succinly-CoA and D-Ribulose 5-phosphate are such two cases, in where rGEM bounds are wider than GEM bounds. Succinly-CoA participates in reaction tetrahydrodipicolinate succinylase with CoA as cofactor pair. Succinly-CoA concentration is tightly constraint due to energetics to synthesize N-Succinyl-2-L-amino-6-oxoheptanedioate, which is an intermediate in L-lysine biosynthesis. The lumped reactions for L-lysine subnetworks do not include this metabolite in the overall stoichiometry, since it is an intermediate. Hence, Succinly-CoA concentration is not constrained in the rGEM. Showing the same behaviour, D-Ribulose 5-phosphate concentration is constrained in arabinose-5-phosphate isomerase reaction, which is in Lipopolysaccharide biosynthesis pathway and producing arabinose-5-phosphate as intermediate.

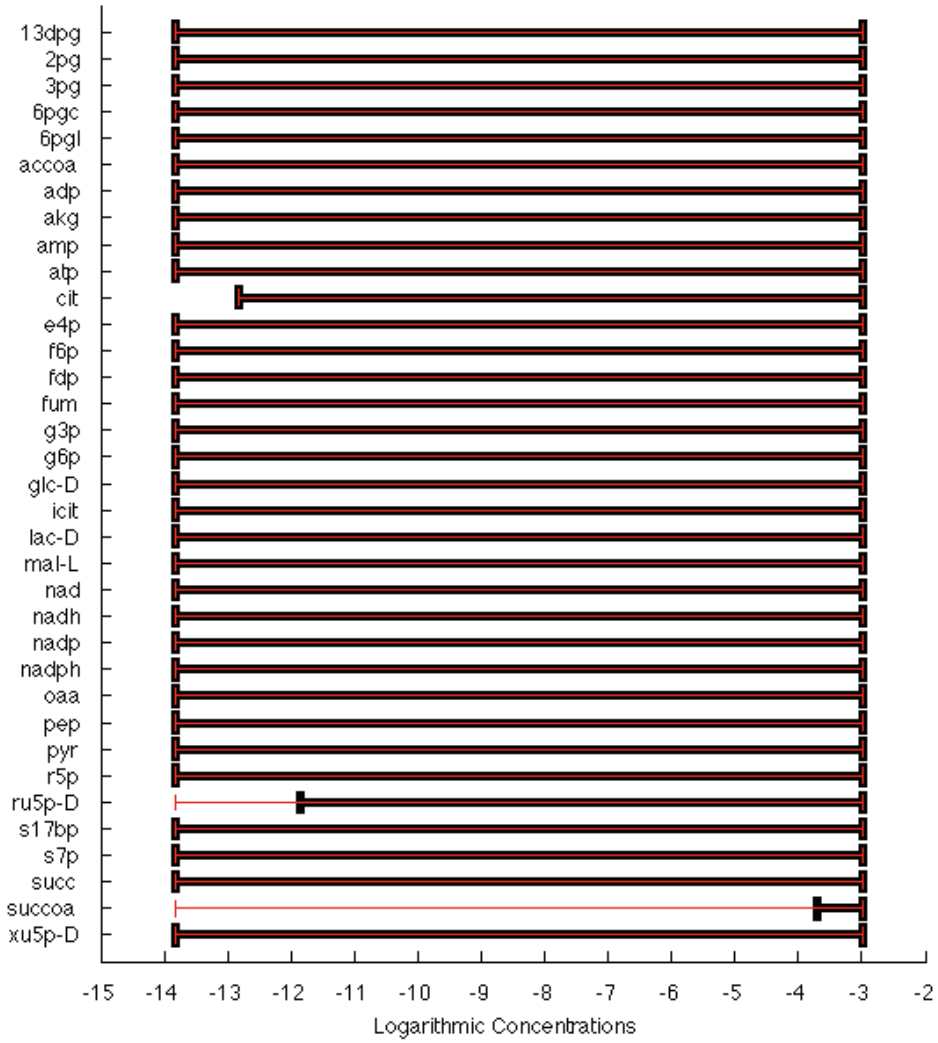

Figure 5-3 CVA on common metabolites between GEM and rGEM (not all shown) Thin red lines are metabolite concentration ranges in rGEM, thick black lines are metabolite concentration ranges in GEM.

5.4 Conclusion

Reduced models have been used for understanding the metabolism for a very long time, long before the first GEM was reconstructed. These top-down models are built to focus on certain parts of metabolism, and built based on the data in the literature. Even though they are far from complete, they proved to be predictive, and helped us understand many aspects of cell metabolism. In this chapter, we merged the invaluable quality that the bottom-up approach brought to network reconstruction with the

reductionist approach that helped researchers to understand the overall behavior of the cells with small scale models. redGEM identifies and builds central carbon metabolism, or any part of the network, by re-defining the term *core*, and by utilizing lumpGEM, it reduces the complexity that emerges from the remote pathways from this core. The output of redGEM is a reduced model that is consistent with its GEM. These reduced models are used in many different areas, such as kinetic modeling, MFA studies, Elementary Flux Modes (EFM) studies and TFA. Moreover, reduced models are promising platforms to compare the central carbon (or any other) metabolism of different species. This helps us to understand the metabolic capabilities and limitations of certain organisms and to investigate the physiological differences that are observed experimentally. In the next chapter, we will focus on this topic and build core carbon metabolisms for different living entities.

Chapter 6 - Comparison of Core Carbon Networks and Biosynthetic Subnetworks of Different Organisms

6.1 Introduction

Central carbon metabolism is one of the highly conserved parts of metabolism among different species, from unicellular bacteria to higher organisms like human[151]. In this definition, core carbon metabolism is defined as the main carbon flows, such as glycolysis, pentose phosphate pathways, TCA cycle, pyruvate metabolism, glyoxylate shunt[21,152]. Even though those pathways exist among different organisms, there is a certain divergence in different enzymes and the way those pathways are optimized[153].

In Chapter 4, we have shown the subnetworks and lumped reactions for an *ad hoc* built core carbon metabolism for *S. cerevisiae,* and identified the differences between *E. coli* and yeast in terms of precursors and cofactor expenditure for the biosynthesis of the same biomass building blocks. The differences between two organisms emerge from mainly 2 sources: *i) the differences along the synthesis pathways, the non-core metabolites that appears on the linear, textbook routes, and how they are balanced in the network, ii) the differences between the core networks*. These distinctions among organism have significant effects on different behaviour of the cells, such as growth yields on the same carbon sources.

In this chapter, we focus on the differences of *core* carbon networks among 4 organisms, *E. coli, P. putida, S. cerevisiae* and human. We chose those 4 organisms to have 2 prokar-

yotic and 2 eukaryotic cells, which have well built GEMs, so that we can minimize the loss of information due to missing information in the models. We used redGEM to extract the core network, and used the same criteria and same subsystems for all 4 organisms to make a fair comparison. Moreover, we generated the subnetworks and lumped reactions for every biomass building block that are defined in the biomass composition in GEMs for D=1 core networks and highlighted the main differences between these 4 organisms.

6.1 Core carbon Network of *P. putida*

P. putida is a gram-negative bacterium, which is in abundance in soil. It is a promising organism for many purposes, from bioremediation to biosynthesis of industrial chemicals. The first GEM of *P. putida* was built in 2008[154], and different GEMs have been reconstructed for different strains of this organism[155,156]. In the latest GEM for *P. putida,* there are 62 subsystems across 2 compartments, periplasm, cytoplasm and extracellular media. The core carbon subsystems are defined in GEM as the following: Glycolysis, Gluconeogenesis, Pentose Phosphate Pathway, TCA Cycle, Pyruvate Metabolism and Oxidative Phosphorylation (Table 6.1).

Table 6-1The Statistics on Starting Subsystems for *P. putida*

Core Subsystems	**# of Reactions**	**# of Tightening Reactions**
TCA Cycle	16	8
Pentose Phosphate Pathway	7	1
Gluconeogenesis	9	6
Glycolysis	21	11
Pyruvate Metabolism	1	0
Oxidative Phosphorylation	14	6
Media Composition	9	-

In total, there are 77 reactions and 81 metabolites in the starting subsystems and 32 tightening reactions. However, tightening reactions are not unique to individual subsystems, and can be shared, and a reaction that is assigned as a different subsystem can be a tightening reaction for another subsystem. Thus, in total, there are 85 reactions, and 81 metabolites in D=0 core network.

Table 6-2 The *core* networks for *P. putida* with different degree of connections D

Degree of Connection	# of Metabolites	# of Reactions
D=1	104	115
D=2	129	138
D=3	154	167
D=4	218	217
D=5	254	259

By performing network expansion with redGEM, we calculated the *core* networks for *P. putida* till degree of connection D=5. The networks are smaller compared to *E. coli* for mainly two reasons: *i) the sizes of GEMs*. *E. coli* GEM has 2585 reactions and 1807 metabolites across periplasm, cytoplasm and extracellular media. On the other hand, *P. putida* has 1053 reactions and 917 metabolites across the two compartments. *E. coli* metabolism has been studies for decades in more details compared to *P. putida* and it is better characterized. As *D* increases, the space that redGEM explores to find pairwise connections between the subsystems is much larger in *E. coli* compared to *P. putida. ii) The anaerobic Electron Transport Chains of E. coli is diverse, and includes many reactions around quinone/quinol pool using many metabolites as substrates and products. P. putida* is a strict aerobe, and cannot grow without oxygen as terminal electron acceptor. However, *E. coli* is a facultative anaerobe and has many anaerobic enzymes in its redox metabolism.

Table 6-3: Comparison between *P. putida* and *E. coli* for alternative subnetworks and unique lumped reactions. The numbers (i:j) represents the statistics for *P. putida* and *E. coli*, respectively. (-) means that these metabolites can be produced in the core network of *E. coli.*

Common *BBB*	Size of S_{min}	# of Alternative S_{min}	# of Lumped Reactions
Glycine	7:6	4:1	2:1
L-Arginine	11:9	3:1	3:1
L-Asparagine	4:1	1:1	1:1
L-Aspartate	2:-	1:-	1:-
L-Cysteine	8:11	1:2	1:2
L-Glutamate	2:-	1:-	1:-
L-Glutamine	3:1	2:1	2:1

L-Histidine	22:17	5:2	5:2
L-Isoleucine	11:10	1:1	1:1
L-Leucine	10:9	2:1	1:1
L-Lysine	12:11	3:1	3:1
L-Methionine	18:20	8:2	6:2
L-Phenylalanine	12:10	1:1	1:1
L-Proline	4:4	2:1	2:1
L-Serine	5:3	1:1	1:1
L-Threonine	7:5	1:1	1:1
L-Tryptophan	19:15	2:1	2:1
L-Tyrosine	12:10	1:1	1:1
L-Valine	6:4	1:1	1:1

Different topology in GEMs and among the *core* networks result in differences in the subnetworks and corresponding lumped reactions generated by lumpGEM for biomass building blocks. We generated the S_{min} subnetworks for *P. putida* (Table 6.3 for amino acids, for the full list, Appendix Table A.6) and performed a comparison with *E. coli*. For 57 BBB, lumpGEM generated 167 subnetworks (1093 for *E. coli*), and 110 unique lumped (246 for *E. coli*) reactions. This multiplicity emerges mainly from big molecules, such as phosphatidylethanolamine, phosphatidylglycerol, peptidoglycan.

6.2 Core carbon Network of *S. cerevisiae*

S. cerevisiae, is one of the mostly studied organisms along with *E. coli*. Some of its properties, such as low pH tolerance make it a potential candidate for many biotechnological purposes[25]. One of the main differences between this yeast and *E. coli* is that *S. cerevisiae* is a eukaryotic cell, and have many compartments. However, many metabolic capabilities between these two organisms are conserved, especially in central carbon metabolism. In this section, we created *core* carbon networks for *S. cerevisiae* and generated subnetworks/lumped reactions for the D=1 *core* carbon network.

For this purpose, we used the GEM of *S. cerevisiae* iMM904 [117], which includes 7 compartments, cytosol, Golgi apparatus, mitochondria, nucleus, Endoplasmic Reticulum, vacuole, peroxisome, and extracellular media. *Core* carbon subsystems of the yeast are named the same as *E. coli*.

iMM904 GEM has 1583 reactions and 1233 metabolites across 7 compartments and extracellular media. In defined D=0 *core* network, there are 95 reactions and 144 metabolites (Table 6.4) including all 6 subsystems.

Table 6-4: Statistics on the Starting Subsystems for *S. cerevisiae*

Core Subsystems	# of Reactions	# of Tightening Reactions
Pyruvate Metabolism	18	12
Citric Acid Cycle	13	10
Pentose Phosphate Pathway	13	3
Glycolysis/Gluconeogenesis	22	7
Oxidative Phosphorylation	19	9
Media Composition	10	-

Similar to *E. coli*, there are many tightening reactions for each subsystem, and some of these tightening reactions are shared among different starting subsystems, resulting in unique 127 reactions and 144 metabolites for D=0 core network.

Following the pairwise connection procedure and tightening analysis for these subsystems, we generated all *core* networks for degree of connection D from 1 to 5 (Table 6.5). The resulting *core* networks are also significantly smaller than *E. coli core* networks, due to the same reason as in *P. putida* case. The size of the GEM of this yeast is smaller compared to iJO1366 reconstruction, and it has less Electron Transport Chains/Oxidative Phosphorylation reactions.

Table 6-5: The *core* networks for iMM904 with different degree of connections

Degree of Connection	# of Metabolites	# of Reactions
D=1	170	180
D=2	181	196
D=3	217	245
D=4	267	321
D=5	302	358

S. cerevisiae is an autotroph, meaning that it can produce all its biomass building blocks from simple sugars or any other electron donors in its surrounding. We employed

lumpGEM on iMM904 to generate all possible S_{min} subnetworks (Table 6.6, for full list, Appendix Table A.7) and corresponding lumped reactions and compared them with *E. coli*.

Table 6-6: Statistics on subnetworks and corresponding unique lumped reactions for amino acids of iMM904. (-) means that these metabolites can be produced in the core network of *E. coli.*

Common *BBB*	Size of S_{min}	# of Alternative S_{min}	# of Lumped Reactions
Glycine	2:6	2:1	2:1
L-Arginine	12:9	2:1	2:1
L-Asparagine	4:1	3:1	3:1
L-Aspartate	2:-	2:-	2:-
L-Cysteine	12:11	6:2	6:2
L-Glutamate	2:-	1:-	1:-
L-Glutamine	2:1	1:1	1:1
L-Histidine	24:17	28:2	28:2
L-Isoleucine	12:10	12:1	12:1
L-Leucine	11:9	2:1	2:1
L-Lysine	10:9	2:1	2:1
L-Methionine	19:20	24:2	24:2
L-Phenylalanine	11:10	1:1	1:1
L-Proline	5:4	2:1	2:1
L-Serine	4:3	3:1	3:1
L-Threonine	7:5	4:1	4:1
L-Tryptophan	18:15	5:1	5:1
L-Tyrosine	11:10	2:1	2:1
L-Valine	6:4	1:1	1:1

The biomass composition of iMM904 is less detailed compared to *E. coli* and has only 44 biomass building blocks, whereas iJO1366 has 102. lumpGEM generated 1093 subnetworks for S_{min} for *E. coli* and 143 for the yeast. However, the number of overall unique lumped reactions for S_{min} subnetworks is very close to each other, 246 and 143 for *E. coli and S. cerevisiae,* respectively. As we discussed in Chapter 4, this signifies that many subnetworks in *E. coli* are equivalent in overall stoichiometry. For iMM904 case, most of the generated subnetworks have different overall stoichiometry, which is a significant difference between those two organisms.

6.3 Core Carbon Network of *Homo sapiens*

The first metabolic network for human metabolism is reconstructed in 2007[122] and updated in 2013[157]. Another human metabolic models were reconstructed by different groups[158,159] Central carbon metabolism, similar to other organisms, plays a very important role in human physiology and the same metabolic subsystems have been conserved in human metabolism (Table 6.7). The *Homo sapiens* GEM, Recon 2 has 7441 reactions and 5063 along 7 compartments, which is almost triple the *E. coli* network. It includes 984 transport reactions between the 7 compartments, and 1601 transport reactions exchanging metabolite with media.

Table 6-7: Statistics on Starting Subsystems for *Homo sapiens*

Core Subsystems	# of Reactions	# of Tightening Reactions
Pyruvate Metabolism	32	20
Citric Acid Cycle	20	42
Pentose Phosphate Pathway	39	20
Glycolysis/Gluconeogenesis	40	47
Oxidative Phosphorylation	10	9
Media Composition	10	-

In starting subsystems of human, there are 203 metabolites and 259 reactions across 4 compartments, cytosol, mitochondria, peroxisome, and endoplasmic reticulum along with extracellular media. The D=0 network for human is significantly bigger than other organisms we have performed the analysis mainly due to tightening reactions. There are many transport reactions between the compartments that include core metabolites but labelled with a different subsystem in the model description.

Table 6-8: Statistics on core networks for *Homo sapiens* with D=1 to D=5

Degree of Connection	# of Metabolites	# of Reactions
D=1	258	432
D=2	351	1183
D=3	454	1411
D=4	637	1753
D=5	734	1906

By using redGEM, we generated the core networks for D=1 to D=5. The most striking result in this network expansion analysis is the degree of freedom that the core networks have. For instance, in *E. coli* D=2 core network, 202 metabolites are participating in 289 reactions, however in Recon 2, with the same degree of connection, 351 metabolites participates in 1183 reactions, representing a very well-connected network (Table 6.8).

After generating the core networks, we applied lumpGEM to generate the subnetworks and lumped reactions for D=1 core network. *Homo sapiens* lacks biosynthesis pathways for many biomass building blocks, including 9 essential amino acids (Table 6.9, for full list Appendix Table A.8). We performed the comparison between *E. coli*, *P. putida* and *S. cerevisiae* under glucose minimal medium, without any additional carbon source, however human cells cannot grow under this minimal medium. Firstly, we employed lumpGEM on Recon2 with the proposed medium in the model. This analysis showed that there are many intermediate metabolites that the cell can uptake from the defined medium to synthesize biomass building blocks. To make a fair comparison, secondly, we tested the capabilities of Recon2 metabolic network to produce biomass building blocks under glucose minimal medium. As expected, 9 essential amino acids cannot be synthesized, and 2 more, L-cysteine and L-Tyrosine. The biosynthesis of L-cysteine requires the uptake of an essential amino acid, L-methionine. With a subnetwork composed of 18 reactions, L-cysteine can be synthesized. The pathway reported for this amino acid in HumanCyc [160] database is composed of 5 reactions, showing the complexity emerging from the non-core metabolites along the biosynthesis pathway. Production of L-tyrosine also requires an uptake of an essential amino acid, L-phenylalanine, and through degradation of this amino acid, L-tyrosine is synthesized. Interestingly, the biosynthesis pathway reported in HumanCyc accounts for the balancing of the non-core metabolites, dihydrobiopterin and tetrahydrobiopterin.

Table 6-9: Statistics on Subnetworks and Lumped Reactions for D=1 *Homo sapiens* rGEM. E stands for *essential.* The amino acids marked with star are not essential, but their synthesis requires an intermediate from the media. (-) means that these metabolites can be produced in the core network of *E. coli.*

Common ***BBB***	Size of S_{min}	# of Alternative S_{min}	# of Lumped Reactions
Glycine	3:6	2:1	2:1
L-Arginine	8:9	32:1	16:1
L-Asparagine	4:1	4:1	4:1
L-Aspartate	2:-	4:-	4:-
L-Cysteine	18*:11	12*:2	12*:2
L-Glutamate	2:-	2:-	2:-
L-Glutamine	2:1	2:1	2:1
L-Histidine	E:17	E:2	E:2
L-Isoleucine	E:10	E:1	E:1
L-Leucine	E:9	E:1	E:1
L-Lysine	E:9	E:1	E:1
L-Methionine	E:20	E:2	E:2
L-Phenylalanine	E:10	E:1	E:1
L-Proline	4:4	6:1	5:1
L-Serine	4:3	2:1	2:1
L-Threonine	E:5	E:1	E:1
L-Tryptophan	E:15	E:1	E:1
L-Tyrosine	2*:10	2*:1	2*:1
L-Valine	E:4	E:1	E:1

6.4 Conclusion

redGEM and lumpGEM allows us to understand the differences in metabolic capabilities between organisms. redGEM defines what is the central carbon metabolism, which fuels the cell and supplies the precursors for biomass building blocks and lumpGEM defines the metabolic capabilities to convert those precursors to biomass building blocks. In this chapter, we used these 2 tools to understand the differences between 4 mentioned organisms and revealed their similarities and difference based on their GEMs. This analysis can be useful in different areas, such as comparison of possible host organisms for targeted chemical production and community modelling. For community modelling, redGEM/lumpGEM provides an excellent platform to compare the metabolic capabilities of the members of the community and to reveal their metabolic interactions, which plays a very important role in the community dynamics.

Chapter 7 - Flux Directionality Profile and Growth Patterns Analysis of *E. coli* under different Carbon Sources

7.1 Introduction

The capacity of a metabolic network to produce precursors for biomass building blocks is one of the major characteristics that determine the growth physiology of the cell. As we have seen in Chapter 4, classical textbooks, such as Neidhardt define certain precursors, which are core carbon metabolites, with certain stoichiometric amounts to produce every component of 1 gDW of *E. coli*. In Chapter 2 and 3, we have built two systematic methods to generate reduced metabolic models that can mimic the Genome Scale model behaviour by defining the overall stoichiometry only through core metabolites. One of the biggest advantages of such reduced models is that they enable us to focus and study on certain part of the metabolism without the complexity, as well as the flexibility that emerges from remote pathways. A well-connected metabolic model is underdetermined since it has more equations (metabolite balances) than variables (reactions), thus having infinite number of possible flux distribution due to the degrees of freedom that the system has. The source of this flexibility can be classified into two main categories, *i) the flexibility in the absolute amount of the flux through a reaction, ii) the directionality of the reaction.* The multiplicity of the former can be studied through different analysis, such as sampling of the fluxes, and further analysis on the generated flux samples like Principle Component Analysis (PCA). However, it is not possible to generate all possible flux distributions in terms of absolute amounts in the solution space, and these methods are usually limited by the capacity of the methods to repre-

sent this space. On the contrary, for the latter, flux directionalities, it is possible to enumerate different possible combinations of flux directionality distributions in a small scale model, since they are free of the multiplicity that emerges from the parts of metabolism that are not under study. Thus, reduced models generated by redGEM/lumpGEM have the potential to be the platform to study different characteristics of metabolic networks with different directionality profiles. In addition, the lumped reactions generated by lumpGEM do not bring any complexity for the directionality analysis since they are fixed to the biomass building block synthesis direction.

In this chapter, we developed a method called Flux Directionality Profile Analysis (FDPA), which enumerates different, if possible all, Flux Directionality Profiles (FDP) for a given metabolic network by utilizing a Mixed-Integer Linear Programming (MILP) algorithm. FDP is a metabolic state of a network that has unique directionalities for each reaction. By imposing these directionalities to the models, FDPA allows us to focus on certain characteristics for each profile, such as biomass production and displacement of reactions from thermodynamic equilibrium. By using FDPA, we generated all possible FDPs of *E. coli* growing under aerobic medium for 2 different conditions, glucose and succinate as the sole carbon sources. Moreover, we ranked the FDPs based on their capability to produce biomass, and revealed couplings between bidirectional reactions under optimal conditions. FDPA also allowed us to study the differences and similarities of metabolic capabilities for the same organisms growing under different carbon sources.

7.2 Materials and Methods

For FDPA, we used the output of redGEM from Chapter 5, a reduced model (rGEM) of *E. coli* iJO1366 growing under aerobic conditions with glucose as the sole carbon source. To be able mimic the GEM behaviour, we kept all unique lumped generated for S_{min} subnetworks.

Firstly, we implied thermodynamics constraints to the model with TFA. In TFA formulation, reactions are split into forward F_{rxn}^{i} and reverse components R_{rxn}^{i}, and for both directions, there are binary variables that control the flux through them in an on/off

manner (FU_{rxn}^{i}, BU_{rxn}^{i}). Thus, every solution that TFA generates is by definition an FDP. Moreover, it is possible to generate alternative FDP with an iterative manner by following these steps:

a. Identify active direction for all reactions for the particular solution.
b. Create a constraint that prevents the solver to choose the same integer combination (integer cuts, FDP)[125] by creating the following constraint:

$FU_{rxn}^{i} + BU_{rxn}^{i+1} + FU_{rxn}^{i+3} + BU_{rxn}^{i+4} \ldots FU_{rxn}^{n} < \sum_{i}^{n} FU_{rxn}^{i} + BU_{rxn}^{i}$

c. Solve the constrained models with an iterative manner till all possible integer cuts are generated.

Following the generation of all FPDs, for every FDP, we tested the biomass production by fixing the directionalities of all reactions obtained from the analysis above. For growth on succinate analysis, we performed exactly the same analysis as growth on glucose case.

7.3 Results and Discussion

The first analysis that we performed on riJO1366 generated is to identify bidirectional reactions (BDRs) for growth under glucose. As a first step, we constrained the model with thermodynamics by TFA formulation, and then performed a Thermodynamic-based Variability Analysis (TVA) to determine these bidirectional reactions. In this network, there are 19 bidirectional reactions, which are distributed among 7 subsystems (Table 7.1). More than one third of the BDRs belong to glycolysis, which states that even though the overall carbon flux in growth under glucose is towards lower glycolysis, the flexibility that the system has allows some of the glycolysis reactions to be able to operate in both directions. Other than glycolysis BDRs, 4 reactions of Pentose Phosphate Pathway (PPP) are bidirectional, and they belong to non-oxidative branch of PPP. There are 3 TCA cycle (Citric Acid Cycle), 3 Pyruvate metabolism BDRs and 1 oxidative phosphorylation BDRs. Interestingly, except the glyoxylate shunt that is composed of only 4 reactions, all the starting subsystems that we have chosen in Chapter 5 as D=0 for redGEM has at least 1 BDR. Oxidative Phosphorylation, which forms the biggest subsys-

tem in the reduced model, has only 1 BDR, indicating the effect of mass and energy constraints on oxidation/reduction processes.

Table 7-1: Bidirectional Reactions in *E. coli riJO1366* grown under glucose as sole carbon source and the subsystems they belong to.

Reaction Names	Reaction Formula	Subsystem
Acetaldehyde Dehydrogenase	ACALD + COA + NAD <=> ACCOA + H + NADH	Pyruvate Metabolism
Acetate Kinase	AC + ATP + H <=> ACTP + ADP	Pyruvate Metabolism
Enolase	2PG <=> H2O + PEP	Glycolysis/Gluconeogenesis
Fructose 6-phosphate Aldolase	F6P <=> DHA + G3P	Glycolysis/Gluconeogenesis
Fructose-bisphosphate Aldolase	FDP <=> DHAP + G3P	Glycolysis/Gluconeogenesis
Fumarase	FUM + H2O <=> MAL-L	Citric Acid Cycle
G3P Dehydrogenase	G3P + NAD + PI <=> 13DPG + NADH	Glycolysis/Gluconeogenesis
Malate Dehydrogenase	MAL-L + NAD <=> H + NADH + OAA	Citric Acid Cycle
Glucose-6-phosphate Isomerase	G6P <=> F6P	Glycolysis/Gluconeogenesis
Phosphoglycerate Kinase	3PG + ATP + H <=> 13DPG + ADP	Glycolysis/Gluconeogenesis
Phosphoglycerate Mutase	2PG <=> 3PG	Glycolysis/Gluconeogenesis
Polyphosphate Kinase	ATP + H + PI <=> ADP + PPI	Oxidative Phosphorylation
Phosphotransacetylase	ACCOA + H + PI <=> ACTP + COA	Pyruvate Metabolism
RU5P-D-3-Epimerase	RU5P-D <=> XU5P-D	Pentose Phosphate Pathway
Succinyl-coa Synthetase	ATP + COA + SUCC <=> ADP + PI + SUCCOA	Citric Acid Cycle
Transaldolase	G3P + S7P <=> E4P + F6P	Pentose Phosphate Pathway
Transketolase1	R5P + XU5P-D <=> G3P + S7P	Pentose Phosphate Pathway
Transketolase2	E4P + XU5P-D <=> F6P + G3P	Pentose Phosphate Pathway
Triose-phosphate Isomerase	DHAP <=> G3P	Glycolysis/Gluconeogenesis

For *E. coli*, with 19 BDRs, FDPA generated 8961 FDPs. Theoretically, 19 bidirectional reactions can generate up to 2^{19} FDPs, which is around ~500 K. However, flux coupling in the metabolic network reduces the possible number of FDPs by not allowing all combinations between BDRs, thus showing that not all combinations between the bidirectional reactions are viable for the organism. This also shows implicitly that for groups of FDPs, some BDRs are fixed to certain directions, and there are some other BDRs that can change sign, however, the behaviour of these FDPs for certain characteristics, such as theoretical maximum growth may be the same. A switch in these indispensible directionalities directly affects the characteristics of the metabolic network, thus switching to a different group of FDPs. In order to test this hypothesis, we tested every FDP with respect to their capacity to produce biomass. To perform this analysis, for every FDP, we

fixed the integer cut that represents the FDP for the model, and maximized through biomass given the mass, thermodynamics and directionality constraints. Then, we sorted the FDPs based on their theoretical maximum yield for biomass accumulation.

As we expected, there are groups of FDPs that have the same theoretical maximum specific growth rate, and there are jumps between the groups of FDPs. A clustering analysis shows that there are 561 different μ_{max} bins that can represent the whole 8961 FDPs (Figure 7.1).

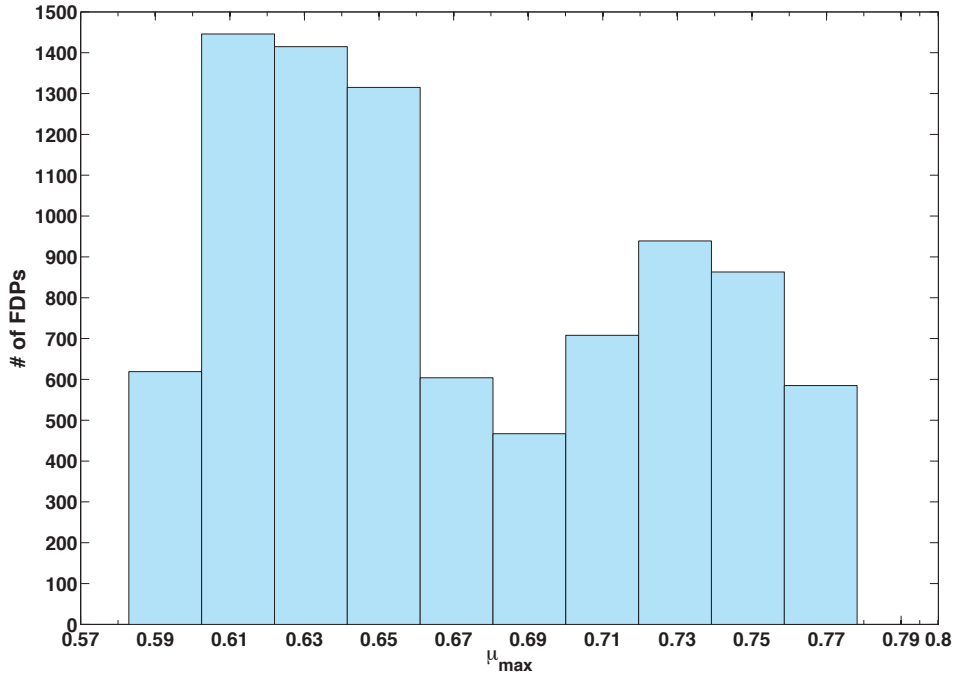

Figure 7-1: The number of FDPs with respect to their theoretical maximum specific growth rate. The glucose uptake is 8.16 mmol/gDWhr for every FDP, and the maximum allowable oxygen uptake rate (OUR) is 20 mmol/gDWhr. The number of FDPs shows a bimodal distribution with the peaks at $\mu_{max} = 0.71 - 0.75$/hr and $\mu_{max} = 0.61 - 0.65$. The experimental growth rate for *E. coli* under the specified conditions is 0.61/hr.

The theoretical maximum specific growth rate of *E. coli* under the specific conditions is 0.7782, and this can be achieved by 36 different FDPs. Among those 36 most efficient FDPs, we see that there are 6 reactions that can switch directions without affecting the

μ_{max}. These reactions are Acetaldehyde Dehydrogenase, Acetate Kinase, Fructose-6-phophate Aldolase, Fructose-biphosphate Aldolase, Phosphotransacetylase and Transaldolase. However, if these 6 BDRs could switch directions without having any coupling between them, then the number of FDPs for this bin should have been 2^6=64. For instance, Acetate Kinase is fully coupled to Phosphotransacetylase, and these reactions are operating in opposite directions, with the directions shown on Table 7.1. Another source of coupling is partial coupling, in other words, directionality coupling, in where the BDRs are coupled to each other in only certain directions, but the coupling is lost when one of the reactions operate in opposite direction. For instance, when Transaldolase reaction operates in Fructose-6-phophate (F6P) direction, Fructose biphosphate Aldolase operates in the direction of Glyceraldehyde-3-phosphate but we don't observe a coupling in the other direction in this yield bin. These 36 FDP indicate that the optimum directionality for glycolysis reactions is towards pyruvate production, as reported in the literature. Moreover, the TCA cycle operates in one, towards CO_2 production, direction without any bifurcation.

The second bin (0.7782/h specific growth rate) is composed of 36 FDPs as well. The minimum hamming distance between the first 36 FDPs and these FDPs is 1. This implies that there is only a single directionality change among the reactions that were set unidirectional in the first 36 FDP to jump to this bin. This reaction is Succinyl-coa Synthetase, and when it catalyses the reaction in succinyl-coa direction, the yield goes down to 0.7756. Again, the same 6 BDRs are generating all the 36 FDPs, similar to the most efficient bin. TCA cycle is still operating in CO_2 direction, even though Succinyl-coa Synthetase belongs to TCA cycle.

The minimum hamming distance between the first and the third bin is also 1, and the BDR that changes its direction in this bin is Transketolase2. The same 6 BDRs can operate bidirectional in this bin as well, however they allow only 24 FDPs. When this non-oxidative PPP enzyme catalyses the reaction in Erytorulose-4-phosphate direction; we observe a drop in growth from 0.7782/h to 0.7748/h. This drop may not be very significant, however combinations of these directionality changes can result in a drop up to 26% in the last bin of FDPs, in where the $\mu_{max} = 0.5829/h$. In this last yield bin, the di-

rectionalities of BDRs are significantly different than the optimum yield bins. The most obvious change is in the direction of glycolysis; glucose-6-phosphate isomerase is operating in glucose-6-phophate direction. However, forcing this reaction in that direction does not have a big effect on growth, and μ_{max} drops to 0.7691/hr. An analysis to identify the minimum number of directionalities to relax to obtain the optimum biomass production indicated 5 reactions, glucose-6-phosphate isomerase, fumarase, triose phosphate isomerase, phosphoglycerate mutase and polyphosphate kinase. We also generated alternative directionality relaxation sets. The second set is composed of 6 reactions, and there are only 5 different alternatives to obtain the maximum biomass yield. This approach can be applied for every yield bin to identify the minimum hamming distance from the optimum bin.

Among 8691 FDPs, there are different profiles that have $\mu_{max} \approx 0.61/h$, which is the experimental specific growth rate for *E. coli* under aerobic conditions with glucose as the sole carbon source. This states that *E. coli* is secreting by-products, and the metabolic network without any directionality constraints (other than pre-assigned directionalities) cannot predict the experimental growth rate, however these FDPs, with a fixed glucose uptake rate, can predict the by-product secretion, thus can be used for different analysis, such as Metabolic Flux Analysis (MFA), in where the experimental data is fitted to a defined metabolic network. The models and the directionalities of reactions for MFA are usually built in *ad hoc* manner and may miss many important links in their network. rGEM models with FDPs that represent the observed growth offer consistent platforms for MFA studies.

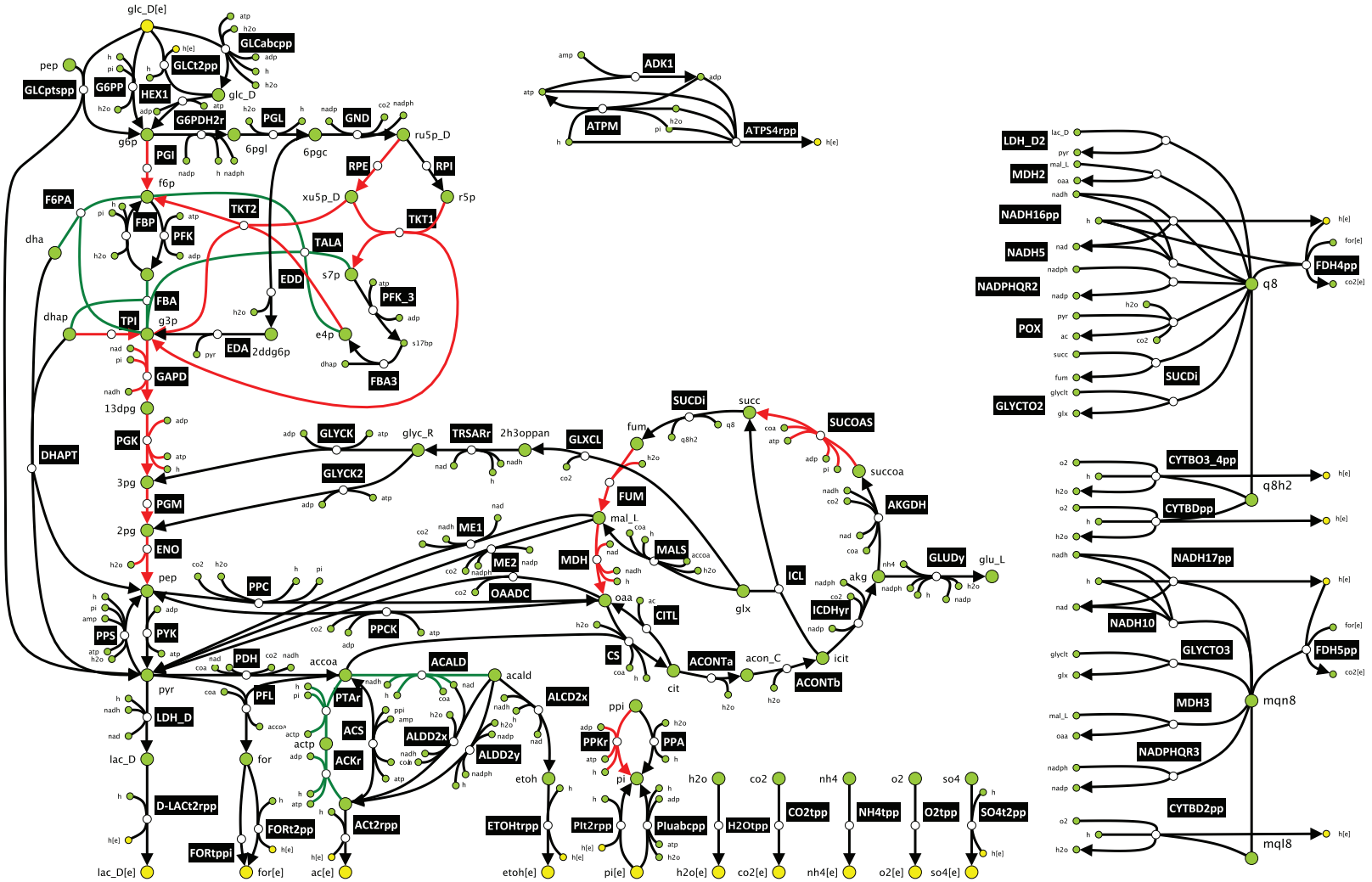

Figure 7-2: Optimum Growth Patterns for Growth under aerobic conditions with glucose as the sole carbon source.

7.3.1 Growth Patters of Aerobic Growth under Succinate as Carbon Source

Growth under different carbon sources or different limiting nutrients requires metabolic switches for the organisms to adapt to the new environment. Expressing different genes, such as genes for transporters or assimilation[161-163], is one of the responses of the organism for this change. Switches in directionality of the metabolic reactions are also important cellular responses to adapt to a different environment. The switches in the directionalities are direct results from the changes in the concentrations of substrates and products. TFA formulation allows us to study these possible changes since it accounts for the steady-state concentration of metabolites. Moreover, studying the growth patterns under different carbon sources can reveal how the optimum patterns differ from one source to another.

The entry point of the carbon source into metabolism plays a very important role for the optimal growth pattern for this carbon source. Most of the core reactions' directions are governed by this fact for optimum growth patterns. As we have seen, the main direction for glycolysis, along with some BDRs, is through pyruvate direction, however this does not mean that this directionality is optimum for growth under different carbon sources. To test this hypothesis, we have generated growth patterns with the same manner as glucose for growth under succinate. The reason for the selection of succinate is to start the metabolism from a remote point from glucose, in this case TCA cycle.

7.3.2 Bidirectionality Analysis for growth under Succinate

Change in carbon source reflects a change in the metabolism, thus the BDRs under different carbon sources are not necessarily the same. In order to make a fair comparison between growths on two different carbon sources, we set the number of carbon atom uptakes as the same both in glucose and succinate case. Since the number of carbon in glucose is 6, we have set the maximum specific succinate uptake as 12.24 mmol/gDWhr (8.16 mmol/gDWhr*6/4). The theoretical maximum specific growth rate for *E. coli* under this environment is 0.583/hr, which is significantly slower compared to growth under glucose (0.7782/hr). Also, the number of bidirectional reactions decreased to 14 (Table 7.2).

Table 7-2: Bidirectional Reactions in *E. coli* iJO1366r grown under glucose as sole carbon source and the subsystems they belong to.

Reaction Names	Reaction Formula	Subsystem
Acetaldehyde Dehydrogenase	ACALD + COA + NAD <=> ACCOA + H + NADH	Pyruvate Metabolism
Acetate Kinase	AC + ATP + H <=> ACTP + ADP	Pyruvate Metabolism
Enolase	2PG <=> H2O + PEP	Glycolysis/Gluconeogenesis
Fructose 6-phosphate Aldolase	F6P <=> DHA + G3P	Glycolysis/Gluconeogenesis
Fructose-bisphosphate Aldolase	FDP <=> DHAP + G3P	Glycolysis/Gluconeogenesis
Malate Dehydrogenase	MAL-L + NAD <=> H + NADH + OAA	Citric Acid Cycle
Phosphoglycerate Mutase	2PG <=> 3PG	Glycolysis/Gluconeogenesis
Polyphosphate Kinase	ATP + H + PI <=> ADP + PPI	Oxidative Phosphorylation
Phosphotransacetylase	ACCOA + H + PI <=> ACTP + COA	Pyruvate Metabolism
RU5P-D-3-Epimerase	RU5P-D <=> XU5P-D	Pentose Phosphate Pathway
Succinyl-coa Synthetase	ATP + COA + SUCC <=> ADP + PI + SUCCOA	Citric Acid Cycle
Transaldolase	G3P + S7P <=> E4P + F6P	Pentose Phosphate Pathway
Transketolase1	R5P + XU5P-D <=> G3P + S7P	Pentose Phosphate Pathway
Transketolase2	E4P + XU5P-D <=> F6P + G3P	Pentose Phosphate Pathway

The first striking observation is that there is no additional BDR to the ones that we observe for glucose case. Moreover, the glycolysis pathway is much more constraint, and 4 out of 8 reactions that were bidirectional under glucose growth becomes unidirectional. 3 reactions in a linear path, phosphoglycerate kinase, glyceraldehyde-3-phosphate dehydrogenase and glucose-6-phosphate isomerase are operating through Glucose-6-phophate direction, and the carbon is carried to Pentose Phosphate Pathway. There is only 1 other reaction in TCA cycle, Fumarase that is fixed to L-malate synthesizing direction. One important observation is that even though the rest 14 BDRs are the same between glucose and succinate case, the allowable flux ranges for these reactions differ under these 2 conditions. The allowable flux bounds for 12 reactions under glucose conditions are encapsulating the ranges under succinate condition, and for all of them the ranges are narrower. For other two reactions, enolase and phosphoglycerate mutase, the ranges do not overlap, however the range for glucose case is still wider compared to succinate case. These observations also suggest that the efficiency of the metabolic network to catabolize the carbon source has a direct effect on the allowable flux ranges, if an organism can grow faster on a certain carbon source compared to another with the same amount of carbon influx, the allowable flux ranges are wider.

7.3.3 Analysis on alternative FDPs and corresponding μ_{max} under Succinate

Following the bidirectionality analysis, we generated all possible FDPs for succinate case. There are 2489 different directionality profiles for *E. coli* that is grown aerobically under succinate as sole carbon source.

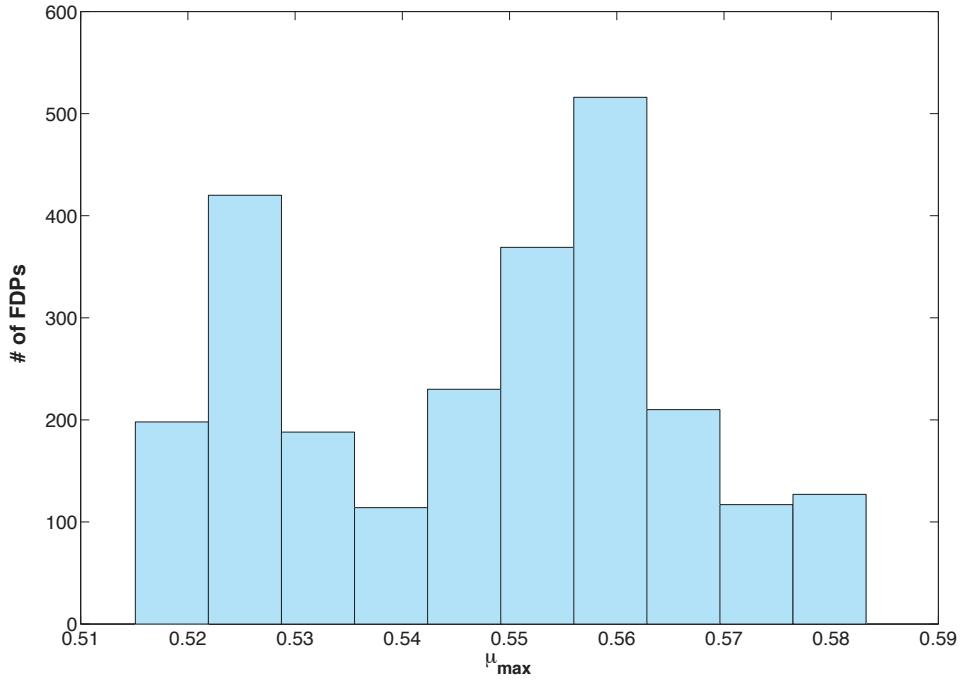

Figure 7-3**:** The number of FDPs with respect to their theoretical maximum specific growth rate. The succinate uptake is 12.24 mmol/gDWhr for every FDP, and the maximum allowable oxygen uptake rate (OUR) is 20 mmol/gDWhr. The number of FDPs shows a bimodal distribution with the peaks at $\mu_{max} = 0.56$/hr and $\mu_{max} = 0.525$/hr.

The theoretical limit for the possible FDPs is 16384 (2^14). Flux couplings are the reason for the decreased number of FDPs. There are 174 yield bins that can span all 2489 FDPs. Unlike glucose case, the theoretical maximum specific growth rate for the optimum (0.5833/hr) and the minimum yield bins (0.5151/hr) are close to each other. To investigate the optimum growth patterns, we studied individually the first 3 bins, and revealed the optimum directionalities for 14 BDRs and the couplings between them in each bin.

There are 9 FDPs that can achieve the highest 0.5833/hr specific growth rate, and there are 9 other FPDs that can achieve 0.5832/hr. In optimal growth pattern analysis, we treated these bins together. Among those 18 FDPs, there are 5 BDRs, namely acetaldehyde dehaydrogenease, acetate kinase, fructose-6-phophate aldolase, fructose biphosphate aldolase, and phophotransacetylase. Among these 5 BDRs, acetate kinase and phophotransacetylase are coupled to each other, and operate in opposite directions with respect to directions reported in Table 7.2. Growth under succinate requires a glycolysis pathway in the direction through glucose-6-phophate for optimality. Along with the 4 reactions that were set unidirectional under succinate growth, the BDRs from glycolysis are set to the same direction in these optimum FDPs. Enolase and phosphoglycerate mutase are operating in D-Glycerate 2-phosphate and D-Glycerate 3-phosphate direction, respectively. Moreover, TCA cycle is operating in the same direction as it operates optimally under glucose condition.

The μ_{max} for the second yield bin in 0.5822, and there are 36 FDPs in this bin. Along with the 5 BDRs from the optimum bin, there is a new BDR for these FDPs, which is transladolase. Transaldolase does not have any coupling with the other BDRs, hence it doubled the number of possible FDPs for this bin. Enolase and phophoglycerate mutase are still coupled to each other in this yield bin. The minimum hamming distance between this bin and the optimum bin is 1, and the reaction responsible for this yield drop is transketolase1. When this reaction operates in glyceraldehyde 3-phosphate direction, the yield drops to 0.5822/hr.

For the third bin, the specific growth rate drops from 0.5822/hr to 0.5811/hr and the minimum hamming distance between the optimum bin and this bin is 3. Transletolase1 is operating in the glyceraldehyde 3-phosphate direction. Moreover, ribulose 5-phosphate 3-epimerase operates in d-xylulose 5-phosphate direction, and transaldolase is fixed in fructose-6-phophate direction. All these 3 enzymes were catalysing the reactions in reverse direction in the optimum bin compared to the third optimum bin.

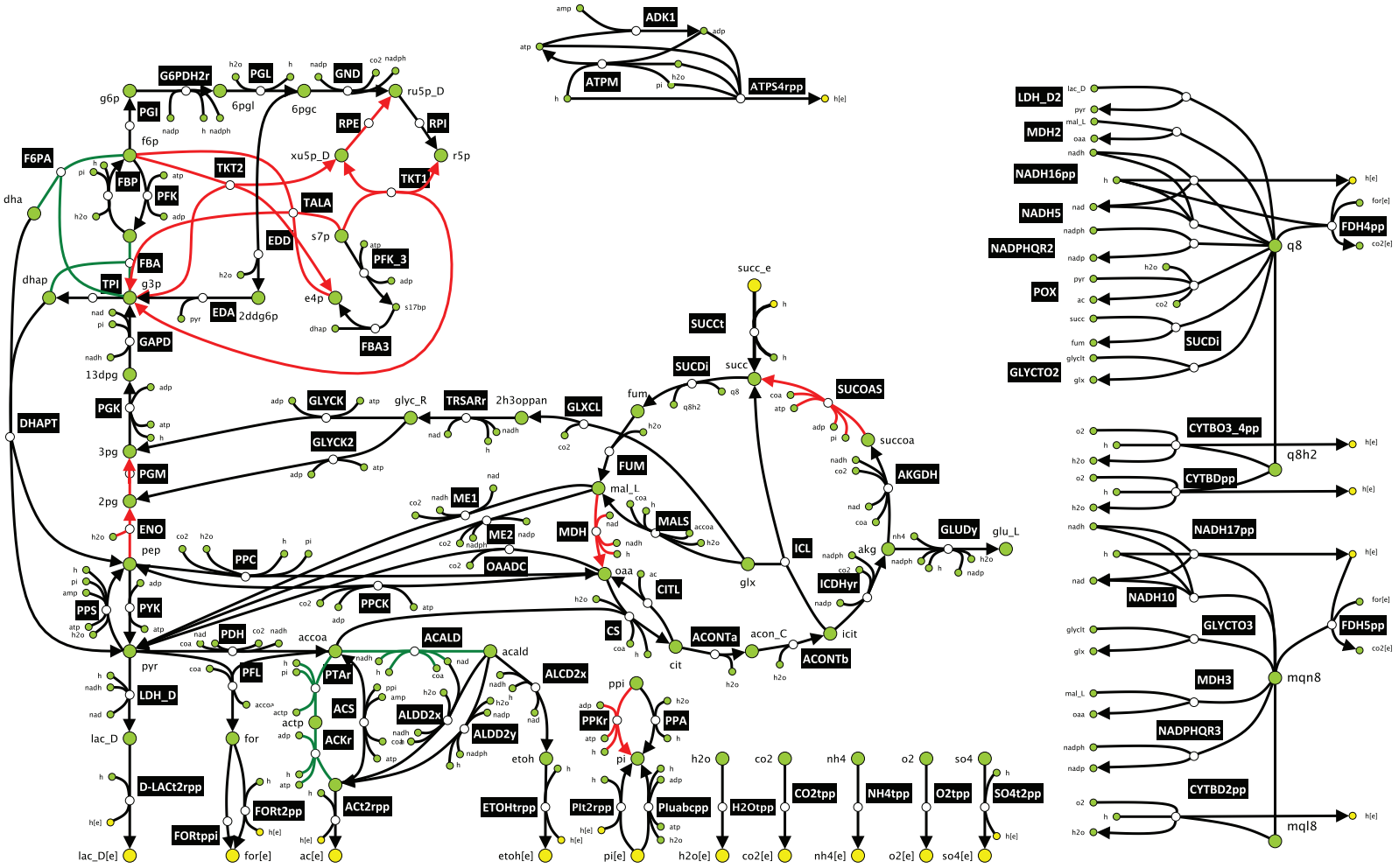

Figure 7-4: Optimum Growth Patterns for Growth under aerobic conditions with succinate as the sole carbon source

Under succinate growth, the oxidative phosphorylation and electron transport reactions are unidirectional, and the directions are the same for both glucose and succinate conditions.

7.4 Conclusion

Metabolic enzymes can catalyse reactions in both directions, unless they are kinetically irreversible. This property allows the cells to adapt to changing environments such as growth media with different carbon sources. Even under the same media, due to the changes in the metabolite concentrations, some reactions can change their directionality. Many reactions in *E. coli* metabolic network have ΔG^0_{rxn} close to 0; hence they are susceptible to directionality switches. In this chapter, we focused on the multiplicity that emerges from bidirectional reactions in metabolic networks. By FDPA, we enumerated all possible directionality profiles for *E. coli* under aerobic conditions, and revealed the differences between growth under glucose and growth under succinate. Slicing the solution space into FDPs gives us the opportunity to study the differences between directionality profiles in terms of biomass yield, target chemical production, displacement from thermodynamic equilibrium. Some of these FDP can explain experimentally observed behaviour of the organisms, and some can fail in this task. Identifying physiologically relevant FDPs helps us to point possible bottlenecks in the metabolic networks for certain tasks, and can reveal the enzymes that controls the system output through kinetic models built specifically for the selected FDPs.

Chapter 8 - *super E. coli*: Exploring Metabolic Capabilities of Central Carbon Metabolism

8.1 Introduction

In the previous 7 chapters, we focused on metabolic networks, mainly GEMs, which are built based on GPR associations. GPR associations, by definition rely on the known biochemistry reported in the databases, and GEMs can only capture the metabolic capabilities based on literature. However, metabolism is a dynamic research area, new enzymes that are catalysing biochemical reactions have been added to the databases each year. For instance, KEGG added hundreds of new biochemical reactions to its database in the update they performed in 2015[164], compared to 2012. In addition, known enzymes are observed to be catalysing new reactions due to enzyme promiscuity. This also brings a complexity for metabolic network analysis, since there can be many possible routes for the carbon flow in the network.

Organisms, such as *E. coli* or *S. cerevisiae* are among the mostly studied organisms in the last century, and their metabolic networks are very well characterized. However, even for these organisms, there are metabolic gaps [165], which means that there are unknown enzymatic activities that are yet to be characterized. In the GEMs of these organism, there are many metabolites that are classified as dead-ends, and there are also computational efforts to predict possible biotransformations acting on these metabolites[166,167]. But these efforts are also limited with the known reactions and they cannot explore novel metabolic capabilities. Along with the mentioned studies, there are also computational efforts to predict possible novel biotransformations[168]. However, these computational tools focus mainly on predicting biosynthesis pathways for industrially relevant chemicals [6363,169-172]. This approach limits their capabilities

to make predictions with a global approach for missing enzymatic activities and unknown metabolic capabilities in the metabolism.

In this chapter, we focused on the potential metabolic capabilities of *E. coli* by utilizing BNICE.ch (Biochemical Network Integrated Computational Explorer) to generate all known and novel reactions around its central carbon metabolism. We determined the effect of the novel reactions on the overall GEM behaviour in terms of biomass yield. Then we reduced the complexity that is emerging from thousands of possible novel biotransformations by identifying smaller sets of novel reactions that can produce the same overall behaviour with an MILP formulation similar to lumpGEM. Finally we have performed a gap filling analysis to determine biotransformations that can connect the dead-end metabolites to native GEM network, thus proposing reactions for the missing enzymatic activities already observed in *E. coli*.

8.2 Materials and Methods

To explore the metabolic capabilities of *E. coli*, we used the BNICE.ch framework[29]. BNICE.ch is framework that is capable of reproducing enzymatic reactions reported in databases (all reactions with a defined biochemistry in KEGG). Moreover it explores and predicts possible novel biotransformations based on the known biochemistry. The underlying idea in BNICE.ch framework is to find patterns in enzymatic reactions and to formula generalized enzyme reaction rules that can capture several similar biochemical reactions. BNICE.ch represents these enzymatic rules by defining the reactive sites of the compounds as two-dimensional molecule fragments as shown in Figure 8.1.

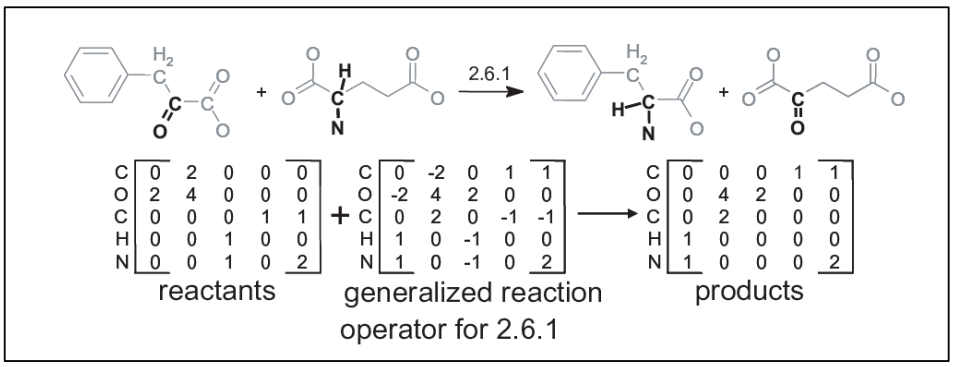

Figure 8-1: Example of Representing the Enzymatic Reaction of phenylpyruvate and glutamate to produce phenylalanine and 2-oxoglutarate. The rule is named as generalized, since in the representation of the reaction, the substrate is not specified, and this rule can catalyze several similar reactions that share the same reactive sites on their substrates [29].

BNICE.ch framework has a database of manually formulated generalized reaction rules that serve to not only to reconstruct known reactions, but also to generate novel enzymatic reactions. These reaction rules are generated with manual curation based on known biochemical reactions in databases, which are classified based on enzyme commission (EC) system [173]. In EC classification system, every enzyme is represented with 4 digits, which count for the biochemistry of the reaction, the reactive site(s) that the enzyme acts on, the cofactor that the enzyme uses and the specific substrate, successively. BNICE.ch rules do not account for the fourth level EC number, since they are not designed as substrate specific. This characteristic is the key point for the generalization of the rules, which is also the key point for designing novel reactions. The rules are defined for unidirectional reactions, so the ones that catalyze the reaction in reverse direction are also created in order to account for the bidirectionality for these reactions. BNICE.ch has two operating functions, *i) forward mode, ii) retrobiosynthesis mode.* In forward mode, BNICE.ch uses all the reaction rules and generates all possible reactions and metabolites from one or a set of starting compounds along with all cofactor pairs defined in BNICE.ch [174]. Retrobiosynthesis mode is mainly used to design linear biosynthesis pathways for the bio-production of target chemicals. For this study, we utilized the forward mode of BNICE.ch to generate a network that includes all central carbon metabolism of *E. coli* and all possible non-native *E. coli* or novel reactions.

8.2.1 Defining the compound space for BNICE.ch

Apart from the known compounds, BNICE.ch can also generate novel, hypothetical metabolites, along with novel reactions, however for this study we limited the scope of results to known compounds. Moreover, even in the known compounds space, there is a complexity that emerges from different databases, such as KEGG or PubChem [175]. BNICE.ch can screen the generated compounds to both these databases, however, to focus on biological compounds space, we constrained the algorithm to generate only biological compounds, and did not include the possible chemical compounds that have never been observed in any biological systems.

8.2.2 Generating the super network of *E. coli*

We created a super network of *E. coli* by merging the generated metabolites and reactions with the GEM of *E. coli*. BNICE.ch reports the generated compounds with their database IDs, hence we could easily map the generated metabolites to the GEM metabolites. Since we do not have any information of the novel reaction directionalities, we integrated them in the network as bidirectional. We also performed Group Contribution Method[40] to estimate Gibbs free energy of formation of the newly added metabolites. Moreover, we estimated the pKa values for these compounds through Marvin software, which can perform pKa estimations based on the 2-D structures of the molecules (Marvin 6.2.3, 2014, ChemAxon (http://www.chemaxon.com)).

8.2.3 Generating yield increasing sets in the super network

To generate the yield increasing sets of novel reactions, we formulated a Mixed-Integer Linear Programming algorithm similar to lumpGEM, to enumerate possible sets that can increase the yield with a minimum of 95% theoretical maximum with the following steps:

a) We created binary use variables $FUSE_i$ for each non-native *E. coli* reaction.
b) We generated a constraint for every non-native *E. coli* reaction that will control the flux through these reactions as:

$$F_{rxn,i} + B_{rxn,i} + C.FUSE_i \leq C$$

C is the number of carbon atoms that the cell uptakes from its surrounding. If the cell can uptake multiple carbon sources, and the number of carbon atoms is not definite, an arbitrary big number can substitute for *C*.

c) We maximize

$$\sum_{i}^{\#\ of\ non-native\ rxns} FUSE_i$$

such that:

$$S.v = 0 \tag{8.1}$$

$$v_{Biomass} \geq .0.95 * \mu_{max} \tag{8.2}$$

d) We generated all alternative solutions exhaustively with an iterative manner by creating the following constraint after every solution.

e)

$$\sum_{k}^{\#\ of\ R^{set_i}} FUSE_{R_k^{sub}} > 0$$

where set_i is the generated set of non-native reactions from Step c.

8.3 Results and discussions

We used the latest GEM of *E. coli* to determine the central carbon metabolism, and selected Glycolysis, Pentose Phosphate Pathway, TCA cycle, pyruvate metabolism, glyoxylate shunt, glycerol synthesis pathway and a few amino acid synthesis pathways that are very close to central metabolism, namely glutamate, aspartate and glutamine. The resulting network is composed of 76 reactions and 67 metabolites (Figure 8.2). All the reactions in this network are in KEGG database, and they have characterized biotransformations, thus can be generated by BNICE.ch. We identified 45 generalized reaction rules that BNICE.ch requires to reproduce the selected 76 central carbon reactions of *E. coli*. This indicates that some rules can reproduce more than 1 reaction for 2 main reasons: *i) enzyme promiscuity in the 4th level EC* and *ii) generalized reaction rules which are not substrate specific.* To account for the possible bidirectionalities in the network, we also selected the reverse reaction rules, thus resulting in 90 generalized reaction rules for further analysis.

As we analyzed in Chapter 4, *E. coli* has all the metabolic pathways for every biomass building block to make 1 gDW of cell, thus can grow under minimal media on single carbon sources. In this chapter, we focus on growth on glucose, and run BNICE.ch on glucose as the only starting compound, along with the cofactor pairs that are coupled to the reaction rules. These cofactor pairs are also determined based on the central carbon

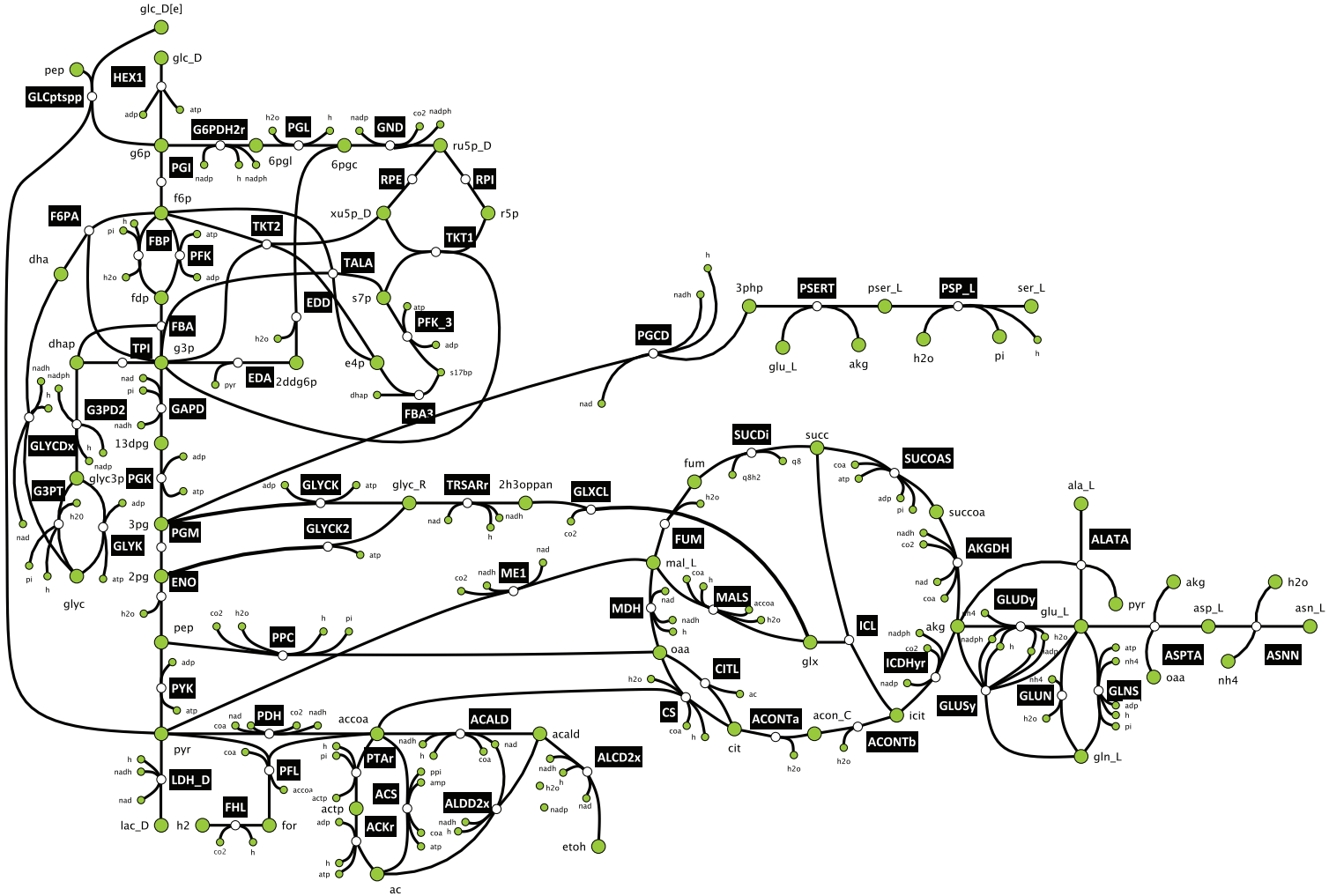

Figure 8-2: The selected network for Super *E. coli* Analysis

metabolism and other cofactor pairs that are not a part of the selected network are excluded from the analysis.

8.3.1 Generation of the *de novo* Network by BNICE.ch

We used the forward mode of BNICE.ch (see Material Methods) and run the algorithm for 16 generations, and after 16th generation, BNICE.ch could not generate any more reactions or metabolites. This convergence indicates that the existing reaction rules cannot generate any more biotransformation between all the generated compounds from the whole 16 generations. The final network is composed of 565 metabolites and 7804 reactions, which forms a well-connected network (Table 8.1). This also shows the promiscuity of the enzymes and the generality of the rules since there are only 90 reaction rules that can potentially catalyse all 7804 reactions. Starting from glucose, there are 12 enzymatic steps to reproduce all 76 reactions of the central carbon metabolism. Moreover, these 90 reaction rules can catalyse 174 more native *E. coli* reactions belonging to subsystems other than central carbon, implying that these rules are also active in other parts of the metabolism. In Chapter 4, we claimed that the existing subsystem definitions in GEMs and databases are not complete, and we re-defined the pathways for biomass building blocks by identifying active enzymes for their biosynthesis. This result is in accordance with this claim from a different perspective, and indicates one more level of connectivity in metabolic networks, the shared biochemistry rules among different parts of the metabolism.

Table 8-1Statistics on the *de novo* Network by BNICE.ch

Generation	Total # of Reactions	# of KEGG Reactions	# of *E. coli* Reactions	Total # of Compounds	# of *E. coli* Compounds
0	0	0	0	19	19
1	18	9	6	28	25
2	80	26	16	45	35
3	197	62	33	83	56
4	553	164	75	159	95
5	1119	322	115	223	130
6	1730	438	143	277	155
7	2419	545	171	334	184
8	3268	648	194	393	210
9	4266	752	222	439	233

10	5247	839	238	478	244
11	6010	892	243	513	251
12	6714	930	247	540	258
13	7441	953	250	555	262
14	7780	965	250	562	262
15	7801	973	250	564	262
16	7804	975	250	565	262

8.3.2 Topology of *de novo* metabolic network of *E. coli*

We then merged the BNICE.ch network with the genome scale network of *E. coli.* Since we build this network around the central carbon metabolism and all those reactions occur in cytoplasm, we introduced every non-native reaction to GEM as cytoplasmic. The super network is composed of 2176 metabolites and 6454 reactions. This signifies a 20.4% increase in number of metabolites and 1.6 times increase in number of reactions. The number of reactions did not increase as much as the number of generated non-native *E. coli* reactions. A closer look into the generated network reveals that there are thousands of same biotransformations with opposite directions. This is due to the forward and backward reaction rules applied by BNICE.ch. Since GEMs account for this bidirectionality through upper and lower bounds for the fluxes, we did not need to introduce these same biotransformations to the super network as 2 separate reactions. We did not put any directionality constraints to the novel reactions in the network, and introduced them as bidirectional for FBA and TFA analysis. Moreover, the reported *E. coli* metabolites in Table 8.1 correspond to all metabolites that have been observed in all strains of *E. coli* in the KEGG database. There are 64 metabolites in KEGG that are reported as *E. coli* compound but not a part of the iJO1366 network.

Firstly, we performed an analysis on the dead-end metabolites of the wild type *E. coli* and the super network. There are 208 dead-end metabolites in the native *E. coli* metabolism. 9 of these metabolites are no longer dead-end in the *de novo* network, and there is more than 1 non-native reaction that includes these metabolites in their stoichiometry. Moreover, all these reactions are novel, suggesting an unknown biotransformation. Generating possible biotransformations based on a certain part of metabo-

lism brings an advantage for our approach over traditional gap filling methods, since gap-filling analysis is performed on individual biomass building blocks. Moreover, these methods do not account for alternative solutions, and does not span the whole possible biotransformations.

Table 8-2**:** List of metabolites that were dead-end metabolites in the native *E. coli* and are not dead-end in the *de novo* metabolic network. The reaction steps that could connect these metabolites to the native metabolites are novel hypothetical reactions generated by BNICE.ch.

Metabolites	# of Reactions	# of Novel Reactions
2,3-Dioxo-L-Gulonate	3	3
2,5-Diketo-D-gluconate	2	2
2-Phosphoglycolate	2	2
4-Hydroxy-L-Threonine	5	5
P1,P5-bis(5-Adenosyl) Pentaphosphate	11	11
1-Aminopropan-2-ol	72	72
1-Deoxy-D-Xylulose	10	10
Gamma-Hydroxybutyrate	15	15
Oxalate	17	17

8.3.3 Characteristics of the *de novo* network, super *E. coli*

The properties of the *de novo* network are significantly different than the native *E. coli* GEM. The theoretical maximum specific growth rate of wild type *E. coli* network is 0.997/hr under aerobic conditions with 10-mmol/gDWhr specific glucose uptake. The *de novo* network, on the other hand can grow 1.46/hr under the same conditions. This indicates a 46% increase in the biomass yield, and it is the maximum achievable yield based on carbon mole glucose over carbon mole biomass, which becomes 1 in this case. In other words, under perfect conditions, the cell does not need to secrete CO_2 to sustain a high growth yield. This behaviour is unexpected, and requires a systematic approach to understand its characteristics.

8.3.4 Sets of reactions that increase the biomass yield

There are 4647 non-native reactions in the super network, and it is not possible to reveal the individual contributions of the novel reactions for this significant increase in

the yield just by optimizing for biomass production. This is mainly due to the degrees of freedom that the system has, plus the complexity that emerges from similar biotransformations among non-native reactions. The first step that we have taken to interpret the effect of non-native reactions and to reduce this complexity is to perform Flux Variability Analysis (FVA). In this approach, we expected to capture the indispensible novel reactions that increase the biomass yield while we imposed a lower bound of 1.46/hr on biomass production. However, FVA indicated that there is not even a single non-native reaction that has a lower allowable absolute flux value bigger than 0. This signifies that the novel reactions increasing the yield can substitute each other, and without an enumeration method, it is not possible to determine the individual or sets of novel reaction that can achieve this goal. To overcome this limitation, and reduce the complexity among the non-native reactions, we formulated an MILP method similar to lumpGEM that was discussed in Chapter 4 (See Materials and Methods) and enumerated all possible sets of those novel reactions. In this analysis, we did not enforce TFA constraints to generate all possible sets and included only mass balance constraints, and then we tested these sets in GEM for thermodynamic feasibility.

Table 8-3**:** Statistics on yield increase sets, FBA/TFA. Total number of generated sets of reactions that increase the yield in FBA, and the number of those that are thermodynamically feasible.

Length	# of generated sets	# of TFA feasible
2	136	7
3	6705	1075
4	4200	1677
5	7	4
Total	11048	2763

Our method generated 11048 sets composed of different lengths, starting from 2 till 5. Surprisingly, 1140 reactions can form all 11048 sets, and 135 of them were KEGG reactions. This corresponds to ~15 percent of the total 7804 reactions proposed by BNICE.ch, which is a significant reduction in the complexity of the super network.

8.3.5 Thermodynamic feasibility of *E. coli* metabolic network with different sets of non-native reactions

After generating all possible yield increasing sets, we tested them for thermodynamics feasibility. In order to perform this analysis, for each set, we created a *de novo* GEM that consists of native *E. coli* metabolism, and the generated set. Among 11048 sets, there were 8285 sets that did not change the biomass yield for *E. coli* with thermodynamics constraints. Moreover, the highest μ_{max} among all sets is 1.39/hr, which is slightly lower than the μ_{max} with only mass balance constraints. There are two main reasons for the drop of biomass yield for the super network. *i) The generated set is thermodynamically infeasible.* In this case, there is at least 1 reaction in the set that cannot carry flux in the desired direction, and this automatically results in no increase in the biomass yield. *ii) the reactions that are coupled to the generated set either cannot carry flux in the desired direction, which will result in no biomass yield increase, or cannot carry flux as high as in FBA case, which will result in a lower biomass yield increase.*

Following the TFA analysis on all the sets, we focused on the individual sets that are increasing yield under thermodynamic constraint. We generated lumped reactions from the sets following a similar method as lumpGEM (See Chapter 4, Materials and Methods) and concluded that some sets are simply producing excess ATP for the super network, and the cell does not need to activate TCA cycle to couple the ATP synthase flux to Electron Transport Chains. This indicates that the limiting factor for *E. coli* for the carbon loss through CO_2 is ATP availability. However, for some other sets, we could not conclude the main reason for the yield increase, and performed a deeper analysis to understand the network behaviour.

In a specific example, by following the carbon flow in the *de novo* network, we observed that a *cycle* is formed, which synthesizes ATP and pumps out protons that are coupled to ATP synthesis. In this cycle, the novel step is the biotransformation of succinate to homoserine, and with 5 successive native reactions, the cycle is closed by the transformation of fumarate to succinate. Another reaction, NADH dehydrogenase, is also coupled to this cycle and pumps out protons from cytosol to periplasm. The second reaction

of the set converts (R)-2-hydroxyglutarate to isocitrate, since (R)-2-hydroxyglutarate is not a native metabolite, and must be balanced in the network. This cycle is very similar to TCA cycle, however there is no CO_2 loss in any step, thus resulting in no carbon loss.

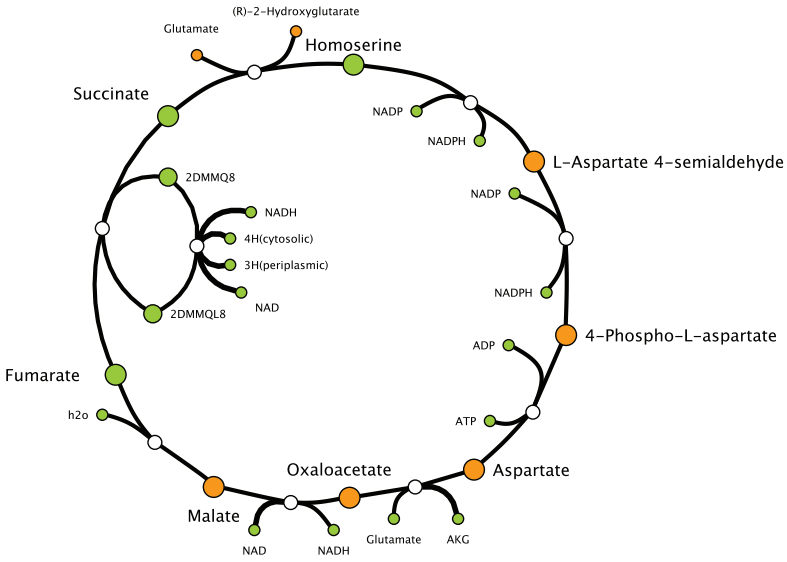

Figure 8-3: An example of sets that increase the yield towards biomass production. The overall reaction for this set is ADP + H^+(c) + NADH + $NADP^+$ + Phosphate <=>ATP + H_2O + NAD^+ + NADPH + H^+(p). This cycle is producing ATP, NADPH and is pumping out protons, which are coupled to ATP synthase reaction.

8.4 Conclusion

Despite the accumulated knowledge in years, we still do not comprehend the full metabolic capabilities of *E. coli*. Using GEMs to identify the metabolic knowledge gaps is a promising approach to determine the missing information. In this chapter, we focused on the central carbon metabolism of *E. coli,* and studied its metabolic potential by using BNICE.ch. BNICE.ch allowed us to investigate the potential non-native reactions that can be catalysed by the enzymatic rules already existing in the organism. Moreover, it suggested reactions acting on the dead-end metabolites in the native network. We extended the scope of the -study by identifying sets of non-native reactions that can increase the yield towards biomass. This study indicates the importance of a global approach to investigate the metabolic capabilities of organisms and its advantages over targeted gap-filling analysis. This method is unbiased since it can be applied to any organism; moreo-

ver, it can be performed on any part of the metabolism. It also indicated the common biochemistry rules among different parts of the metabolic network, thus revealing one more layer of connectivity between subsystems.

Chapter 9 - Conclusions and Perspectives

In this thesis, we had mainly 2 perspectives, *i) reducing the complexity of metabolic networks* and *ii) characterization of metabolic networks through the methods that we have developed for reducing this complexity.* This complexity emerges from different characteristics of metabolic networks. In the first seven chapters, we focused on 2 sources: *i) the size of the metabolic networks* and *ii) underdetermined nature of the metabolic networks, and the large degrees of freedom that the systems have.* In Chapter 8, we focused on the complexity of the metabolic capabilities of organisms and proposed methods to characterize the metabolic potentials of the enzymatic capabilities of *E. coli.*

Bio-thermodynamics that is described in details in Chapter 1 is one of the most popular methods to reduce the solution space for metabolic networks and to determine directionality of reactions that have not been observed as catalytically irreversible. By discussing different approaches, we have reviewed the existing state of the methods; in addition, we have discussed and revealed other potential areas that bio-thermodynamics can play an important role, such as plant metabolism in where energy metabolism is playing a very crucial role. For future studies, merging the computational chemistry methods with existing bio-thermodynamics methods is a promising approach for increasing the coverage of thermodynamic properties and for further improvement of thermodynamic estimations. Moreover, this approach will make it possible to study on the organisms that survive under extreme conditions, extremophiles, since the thermodynamic estimations can be made for different temperatures and pressures through high-level quantum chemical calculation.

In Chapter 3, we focused on a new method by utilizing bio-thermodynamics with a systems biology approach through TFA. We showed that bio-thermodynamics is not only

useful for assigning directionalities and determining allowable concentration ranges for metabolites thus distance from thermodynamic equilibrium, but also useful to understand the overall system behaviour, in this specific case Gibbs free energy dissipation and efficiency of electron transport chains and oxidative phosphorylation. This method is applicable to any genome scale model, which makes it a powerful approach to analyse the relation between thermodynamic characteristics and the observed physiology of the organisms. For future studies, this method can be applied to other organisms to understand the growth characteristics, by-product formation and their coupling with engineered pathways. Moreover, Gibbs free energy dissipation can be used as a criterion in engineering organisms for certain tasks, such as biosynthesis of target chemicals. Possible non-native synthesis pathways can be ranked according to their additional dissipation burden for the host organism.

In Chapter 4 and 5, we focused on reducing the complexity on a topological level, and by lumpGEM, we reduced the GEM for the synthesis of individual biomass building blocks, and we re-defined the biological pathway definition by showing that the active enzymes for the synthesis of a certain metabolite form a nested subnetwork, rather than linear as shown in databases, such as KEGG and EcoCyc. This approach led to the development of core reduced models around the parts of metabolism under interest by redGEM in Chapter 5. redGEM generated representative models from GEM by re-defining the central carbon metabolism, and by reducing the stoichiometry through lumpGEM. The consistency between rGEMs and GEMs makes these reduced models powerful platforms for different studies that focus on certain parts of the metabolism. Moreover, this approach is very promising for many other applications, such as community modelling, in where microbial consortia with many organisms are under study. Revealing the dynamics between the individual members of the consortia is crucial to understand the community behaviour, growing in different environments, from soil to clinical cases, for instance burnt skin. The sizes of these community models are manageable for studies, such as FBA, TFA but the complexity of the network and uncertainty in the kinetic parameters makes it very difficult to build representative kinetic models.

redGEM and lumpGEM also have the potential to uncover the differences between the central carbon metabolism and to indicate the different synthesis routes that evolved for the same biomass building blocks among different organisms. In Chapter 6, we focused on these differences, and showed the characteristics of different central carbon networks among prokaryotic and eukaryotic cells. This study has the potential to identify some of the evolutionary differences between the strains of the same species or tissue specific cells in higher organisms, such as mouse or human.

In Chapter 7, we enumerated all the possible Flux Directionality Profiles for *E. coli* under different conditions by using the generated rGEMs from Chapter 5. This approach can be used for Metabolic Flux Analysis, in where the experimental data can be fitted to different directionality profiles. As a future study, growth patterns analysis can reveal minimum hamming distances between different yield bins that have been discussed in details in the chapter. Growth pattern analysis also reveals certain directionality profiles with a theoretical maximum yield that is experimentally observed. Moreover, with the same approach, FDPA can be used to study sub-optimal growth conditions, since it also proposes FDPs with lower theoretical maximum growth rate compared to optimally grown conditions. FDPA also has the potential to be merged with different approaches in this thesis, such as P:O Ratio and bioenergetics analysis in Chapter 3. For future work, analysis of directionality profiles with the optimum P:O Ratio and minimum Gibbs free energy dissipation can be applied for the study of the flexibility of organisms under optimum growth conditions.

Enzyme promiscuity and uncharacterized enzymatic activities bring a different type of flexibility and complexity to the metabolic network analysis studies. In Chapter 8, we used BNICE.ch to explore the metabolic capabilities of *E. coli* with known and possible biotransformations by creating a *super* network based on its known biochemistry. This type of analysis has many potential applications, such as increasing the yield through biomass production. Moreover, this approach can be used to analyse the metabolic capabilities of *super E. coli* network for different purposes, such as the biosynthesis of industrially relevant chemicals and provides guidance for synthetic biology and metabolic engineering. Reducing the complexity among thousands of novel reactions also creates

the chance to enumerate different small sets of novel reactions to achieve the metabolic goals listed above. We can then use another tool developed in LCSB, BrigdIT, which makes use of the similarity between reactants and products of the novel reaction and all known KEGG reactions, and based on this similarity, ranks the possible enzymes that can potentially catalyse this novel reaction. Such analysis has the potential to be applied in experimental set-ups through determining potential enzymes to be engineered using protein engineering and directed evolution.

Computational methods are indispensible tools to improve our understanding of cell metabolism and its potential, to guide the experimental set-ups and improve the performance of the organisms for certain tasks. In this thesis, we highlighted different computational efforts to fulfil these goals, and discussed our contribution to the field. By building methods to manage the complexity of the metabolic networks, we proposed novel approaches to characterize cellular metabolism. We also aim to re-think and re-define some established definitions in the field, such as subsystems and metabolic pathways. The methods proposed in this thesis have the potential to be applied to any metabolic network, thus making them unbiased tools for many different studies on metabolism.

Appendix

Table A.1: Reactions that contributes most to Gibbs Free Energy Dissipation of *E. coli*

Reactions	Enzymatic Reaction Formula	$\Delta G^{0'}_{rxn}$
AKGDH	AKG + COA + NAD <=> CO2 + NADH + SUCCOA	-11.69
ATPM	ATP + H2O <=> ADP + H + PI	-7.76
ATPsynthase	ADP + PI + 4 H_P <=> ATP + 3 H + H2O	-8.80
CS	ACCOA + H2O + OAA <=> CIT + COA + H	-8.78
CYTbd2	2 H + MQL8 + 0.5 O2 <=> H2O + MQN8 + 2 H_P	-14.95
CYTbd1	2 H + 0.5 O2 + Q8H2 <=> H2O + Q8 + 2 H_P	-28.67
CYTbo3	4 H + 0.5 O2 + Q8H2 <=> H2O + Q8 + 4 H_P	-20.38
FACOAE60	H2O + HXCOA <=> COA + H + HXA	-4.83
FDH4pp	2 H + Q8 + FOR_P <=> Q8H2 + CO2_P + H_P	-10.48
FDH5pp	2 H + MQN8 + FOR_P <=> MQL8 + CO2_P + H_P	-24.20
FEROpp	4 FE2_P + 4 H_P + O2_P <=> 4 FE3_P + 2 H2O_P	-8.12
FOMETRi	5FTHF + H <=> H2O + METHF	-5.92
FRD3	2DMMQL8 + FUM <=> 2DMMQ8 + SUCC	5.31
G3PD5	GLYC3P + Q8 <=> DHAP + Q8H2	-15.11
G3PD6	GLYC3P + MQN8 <=> DHAP + MQL8	-28.83
G3PD7	2DMMQ8 + GLYC3P <=> 2DMMQL8 + DHAP	-17.98
G6PDH2r	G6P + NADP <=> 6PGL + H + NADPH	-3.39
GAPD	G3P + NAD + PI <=> 13DPG + NADH	-2.15
GLUDy	GLU-L + H2O + NADP <=> AKG + H + NADPH + NH4	11.23
GLYAT	ACCOA + GLY <=> 2AOBUT + COA	7.25
GLYCLTDx	GLX + H + NADH <=> GLYCLT + NAD	-6.21
GLYCLTDy	GLX + H + NADPH <=> GLYCLT + NADP	-5.82
GLYCTO2	GLYCLT + Q8 <=> GLX + Q8H2	-9.76
GLYCTO3	GLYCLT + MQN8 <=> GLX + MQL8	-23.48
GLYCTO4	2DMMQ8 + GLYCLT <=> 2DMMQL8 + GLX	-12.63
HEXt2rpp	H_P + HXA_P <=> H + HXA	1.26
HPYRRx	H + HPYR + NADH <=> GLYC-R + NAD	-4.75
HPYRRy	H + HPYR + NADPH <=> GLYC-R + NADP	-4.36
LDH_D	LAC-D + NAD <=> H + NADH + PYR	4.75
LDH_D2	LAC-D + Q8 <=> PYR + Q8H2	-11.22
MDH	MAL-L + NAD <=> H + NADH + OAA	4.75
MDH2	MAL-L + Q8 <=> OAA + Q8H2	-11.22

MDH3	MAL-L + MQN8 <=> MQL8 + OAA	-24.94
MOX	MAL-L + O2 <=> H2O2 + OAA	-25.49
NADH10	H + MQN8 + NADH <=> MQL8 + NAD	-29.69
NADH17pp	4 H + MQN8 + NADH <=> MQL8 + NAD + 3 H_P	-13.13
NADH18pp	2DMMQ8 + 4 H + NADH <=> 2DMMQL8 + NAD + 3 H_P	-2.28
NADH5	H + NADH + Q8 <=> NAD + Q8H2	-15.97
NADH9	2DMMQ8 + H + NADH <=> 2DMMQL8 + NAD	-18.84
NADPHQR2	H + NADPH + Q8 <=> NADP + Q8H2	-15.58
NADPHQR3	H + MQN8 + NADPH <=> MQL8 + NADP	-29.30
NADPHQR4	2DMMQ8 + H + NADPH <=> 2DMMQL8 + NADP	-18.45
PDH	COA + NAD + PYR <=> ACCOA + CO2 + NADH	-10.53
PFK	ATP + F6P <=> ADP + FDP	-7.68
PFK_3	ATP + S7P <=> ADP + S17BP	-7.26
PGL	6PGL + H2O <=> 6PGC + H	-7.21
PPAKr	ADP + PPAP <=> ATP + H + PPA	-1.15
PPCSCT	PPCOA + SUCC <=> PPA + SUCCOA	-1.16
PPKr	ATP + H + PI <=> ADP + PPI	-1.90
PTA2	H + PI + PPCOA <=> COA + PPAP	4.08
PTAr	ACCOA + H + PI <=> ACTP + COA	7.12
PYAM5PO	H2O + O2 + PYAM5P <=> H2O2 + NH4 + PYDX5P	-18.10
SUCDi	Q8 + SUCC <=> FUM + Q8H2	-2.44
SUCOAS	ATP + COA + SUCC <=> ADP + PI + SUCCOA	-4.09
THD2	NADH + NADP + 2 H_P <=> 2 H + NAD + NADPH	-8.67

Table A.2: Statistics of Subnetworks and Lumped Reactions of *E. coli* for *ad hoc* Core Network

Biomass Building Blocks	**Size of Subnetwork**	**# of Alternative Subnetwork**	**# of Unique Lumped Reactions**
10-Formyltetrahydrofolate	46	24	7
2-Demethylmenaquinol 8	31	24	4
[2Fe-2S] iron-sulfur cluster	19	40	16
[4Fe-4S] iron-sulfur cluster	21	48	16
5-Methyltetrahydrofolate	46	36	4
Acetyl-CoA	48	36	3
Adenosylcobalamin	29	48	15
L-Alanine	2	1	1
S-Adenosyl-L-methionine	43	36	4
L-Arginine	13	2	2
L-Asparagine	5	2	2
L-Aspartate	2	1	1
bis-molybdopterin guanine dinucleotide	46	24	8
Biotin	67	72	8
Chorismate	7	1	1
Coenzyme A	48	35	3
CTP	15	12	9
L-Cysteine	15	6	2
dATP	29	48	16
dCTP	17	24	18
dGTP	30	56	14
dTTP	23	18	12
Enterochelin	19	4	2
Flavin adenine dinucleotide oxidized	39	24	7
L-Glutamine	2	2	2
L-Glutamate	1	1	1
Glycine	8	2	2
glycogen	2	2	2
Reduced glutathione	21	6	2
GTP	28	24	7
Heme O	27	24	8
L-Histidine	21	12	3
L-Isoleucine	12	1	1
L-Leucine	10	1	1
lipoate (protein bound)	40	48	16
L-Lysine	11	1	1
Malonyl-CoA	49	36	3

L-Methionine	25	6	2
5,10-Methylenetetrahydrofolate	45	36	4
molybdopterin cytosine dinucleotide	53	144	48
molybdopterin guanine dinucleotide	44	24	8
Menaquinol 8	44	21	4
Nicotinamide adenine dinucleotide	32	72	21
Nicotinamide adenine dinucleotide - reduced	32	72	21
Nicotinamide adenine dinucleotide phosphate	33	72	21
Nicotinamide adenine dinucleotide phosphate - reduced	33	70	21
phosphatidylethanolamine (dihexadecanoyl, n-C16:0)	44	245	16
phosphatidylethanolamine (dihexadec-9enoyl, n-C16:1)	44	109	14
phosphatidylethanolamine (dioctadec-11-enoyl, n-C18:1)	48	256	16
Phosphatidylglycerol (dihexadecanoyl, n-C16:0)	40	256	16
Phosphatidylglycerol (dihexadec-9-enoyl, n-C16:1)	40	127	14
Phosphatidylglycerol (dioctadec-11-enoyl, n-C18:1)	44	242	16
L-Phenylalanine	11	1	1
Protoheme	13	4	4
L-Proline	5	1	1
Putrescine	7	1	1
Pyridoxal 5-phosphate	7	1	1
Ubiquinol-8	44	85	16
Riboflavin	35	24	7
L-Serine	4	1	1
Siroheme	24	2	2
Spermidine	37	6	2
Succinyl-CoA	48	36	3
5,6,7,8-Tetrahydrofolate	46	24	7
Thiamine diphosphate	46	12	4
L-Threonine	7	1	1
L-Tryptophan	17	2	2
L-Tyrosine	11	1	1
Undecaprenyl diphosphate	15	12	4
UTP	13	6	6
L-Valine	5	1	1
core oligosaccharide lipid A	75	128	24
cardiolipin (tetrahexadecanoyl, n-C16:0)	41	512	32
cardiolipin (tetrahexadec-9-enoyl, n-C16:1)	41	255	28
cardiolipin (tetraoctadec-11-enoyl, n-C18:1)	45	512	32
two linked disacharide tripeptide murein units (uncrosslinked, middle of chain)	32	4	4
two disacharide linked murein units, tripeptide crosslinked tetrapeptide (A2pm->D-ala) (middle of chain)	33	4	4
two linked disacharide tetrapeptide murein units (uncross-	33	4	4

linked, middle of chain)			
two disacharide linked murein units, tetrapeptide corsslinked tetrapeptide (A2pm->D-ala) (middle of chain)	33	4	4
three disacharide linked murein units (tetrapeptide crosslinked tetrapeptide (A2pm->D-ala) & tetrapeptide corsslinked tetrapeptide (A2pm->D-ala)) (middle of chain)	33	4	4
phosphatidylethanolamine (dihexadecanoyl, n-C16:0)	44	256	16
phosphatidylethanolamine (dihexadec-9enoyl, n-C16:1)	44	128	14
phosphatidylethanolamine (dioctadec-11-enoyl, n-C18:1)	48	234	16
Phosphatidylglycerol (dihexadecanoyl, n-C16:0)	40	512	32
Phosphatidylglycerol (dihexadec-9-enoyl, n-C16:1)	40	247	28
Phosphatidylglycerol (dioctadec-11-enoyl, n-C18:1)	44	512	32
Growth Associated Maintenance	27	24	8

Table A.3: Core Carbon Network of *E. coli* for D=0

Reactions	Reaction Formulas	Reactions	Reaction Formulas
DM_ac_e	ac_e <=>	GLYCK2	atp + glyc-R <=> 2pg + adp
DM_akg_e	akg_e <=>	GLYCLTDx	glx + h + nadh <=> glyclt + nad
DMo2_e	co2_e <=>	GLYCLTDy	glx + h + nadph <=> glyclt + nadp
DM_etoh_e	etoh_e <=>	GLYCTO2	glyclt + q8 <=> glx + q8h2
DM_for_e	for_e <=>	GLYCTO3	glyclt + mqn8 <=> glx + mql8
DM_glyc_e	glyc_e <=>	GLYCTO4	2dmmq8 + glyclt <=> 2dmmql8 + glx
DM_glycogenn1	<=> glycogenn1	GND	6pgc + nadp <=> co2 + nadph + ru5p-D
DM_bglycogenn1	<=> bglycogenn1	H2Otpp	h2o_p <=> h2o
DM_lac-D_e	lac-D_e <=>	HEX1	atp + glc-D <=> adp + g6p
DM_mal-L_e	mal-L_e <=>	HYD1pp	2 h + h2 + q8 <=> q8h2 + 2 h_p
DM_pyr_e	pyr_e <=>	HYD2pp	2 h + h2 + mqn8 <=> mql8 + 2 h_p
DM_succ_e	succ_e <=>	HYD3pp	2dmmq8 + 2 h + h2 <=> 2dmmql8 + 2 h_p
ACALD	acald + coa + nad <=> accoa + h + nadh	ICDHyr	icit + nadp <=> akg + co2 + nadph
ACKr	ac + atp + h <=> actp + adp	ICL	icit <=> glx + succ
ACONTa	cit <=> acon-C + h2o	L-LACD2	lac-L + q8 <=> pyr + q8h2
ACONTb	acon-C + h2o <=> icit	L-LACD3	lac-L + mqn8 <=> mql8 + pyr
ACS	ac + atp + coa + h <=> accoa + amp + ppi	LDH_D	lac-D + nad <=> h + nadh + pyr
ADK1	amp + atp <=> 2 adp	LDH_D2	lac-D + q8 <=> pyr + q8h2
AKGDH	akg + coa + nad <=> co2 + nadh + suc-coa	MALS	accoa + glx + h2o <=> coa + h + mal-L
ALCD2x	etoh + nad <=> acald + h + nadh	MDH	mal-L + nad <=> h + nadh + oaa
ALDD2x	acald + h2o + nad <=> ac + 2 h + nadh	MDH2	mal-L + q8 <=> oaa + q8h2
ALDD2y	acald + h2o + nadp <=> ac + 2 h + nadph	MDH3	mal-L + mqn8 <=> mql8 + oaa
ASPO3	asp-L + q8 <=> h + iasp + q8h2	ME1	mal-L + nad <=> co2 + nadh + pyr
ASPO4	asp-L + mqn8 <=> h + iasp + mql8	ME2	mal-L + nadp <=> co2 + nadph + pyr
ASPO5	asp-L + fum <=> h + iasp + succ	MOX	mal-L + o2 <=> h2o2 + oaa
ASPO6	asp-L + o2 <=> h + h2o2 + iasp	NADH10	h + mqn8 + nadh <=> mql8 + nad
ASPT	asp-L <=> fum + nh4	NADH16pp	4 h + nadh + q8 <=> nad + q8h2 + 3 h_p
ATPM	atp + h2o <=> adp + h + pi	NADH17pp	4 h + mqn8 + nadh <=> mql8 + nad + 3 h_p

ATPS4rpp	adp + pi + 4 h_p <=> atp + 3 h + h2o
CAT	2 h2o2 <=> 2 h2o + o2
CITL	cit <=> ac + oaa
CO2tex	co2_e <=> co2_p
CO2tpp	co2_p <=> co2
CRNBTCT	bbtcoa + crn <=> crncoa + gbbtn
CRNCAL2	atp + coa + crn <=> adp + crncoa + pi
CRNCAR	crncoa <=> crnDcoa
CRNCBCT	crn + ctbtcoa <=> crncoa + ctbt
CRNCDH	crncoa <=> ctbtcoa + h2o
CRNDCAL2	atp + coa + crn-D <=> adp + crnDcoa + pi
CS	accoa + h2o + oaa <=> cit + coa + h
CTBTCAL2	atp + coa + ctbt <=> adp + ctbtcoa + pi
CYTBD2pp	2 h + mql8 + 0.5 o2 <=> h2o + mqn8 + 2 h_p
CYTBDpp	2 h + 0.5 o2 + q8h2 <=> h2o + q8 + 2 h_p
CYTBO3_4pp	4 h + 0.5 o2 + q8h2 <=> h2o + q8 + 4 h_p
DHAPT	dha + pep <=> dhap + pyr
DHORD2	dhor-S + q8 <=> orot + q8h2
DHORD5	dhor-S + mqn8 <=> mql8 + orot
DHORDfum	dhor-S + fum <=> orot + succ
DMSOR1	dmso + mql8 <=> dms + h2o + mqn8
DMSOR1pp	mql8 + dmso_p <=> mqn8 + dms_p + h2o_p
DMSOR2	2dmmql8 + dmso <=> 2dmmq8 + dms + h2o
DMSOR2pp	2dmmql8 + dmso_p <=> 2dmmq8 + dms_p + h2o_p
DMSOtpp	dmso_p <=> dmso
DSBAO1	q8 + dsbard_p <=> q8h2 + dsbaox_p
NADH18pp	2dmmq8 + 4 h + nadh <=> 2dmmql8 + nad + 3 h_p
NADH5	h + nadh + q8 <=> nad + q8h2
NADH9	2dmmq8 + h + nadh <=> 2dmmql8 + nad
NADK	atp + nad <=> adp + nadp
NADPHQR2	h + nadph + q8 <=> nadp + q8h2
NADPHQR3	h + mqn8 + nadph <=> mql8 + nadp
NADPHQR4	2dmmq8 + h + nadph <=> 2dmmql8 + nadp
NADPPPS	h2o + nadp <=> h + nad + pi
NADTRHD	nad + nadph <=> nadh + nadp
NH4tpp	nh4_p <=> nh4
NO2t2rpp	h_p + no2_p <=> h + no2
NO3R1bpp	q8h2 + no3_p <=> q8 + h2o_p + no2_p
NO3R1pp	2 h + no3 + q8h2 <=> h2o + no2 + q8 + 2 h_p
NO3R2bpp	mql8 + no3_p <=> mqn8 + h2o_p + no2_p
NO3R2pp	2 h + mql8 + no3 <=> h2o + mqn8 + no2 + 2 h_p
NO3t7pp	no2 + no3_p <=> no3 + no2_p
NTP1	atp + h2o <=> adp + h + pi
NTPP6	atp + h2o <=> amp + ppi
NTRIR2x	5 h + 3 nadh + no2 <=> 2 h2o + 3 nad + nh4
NTRIR3pp	3 q8h2 + 2 h_p + no2_p <=> 3 q8 + 2 h2o_p + nh4_p
NTRIR4pp	3 mql8 + 2 h_p + no2_p <=> 3 mqn8 + 2 h2o_p + nh4_p
OAADC	h + oaa <=> co2 + pyr
PDH	coa + nad + pyr <=> accoa + co2 + nadh
PFK	atp + f6p <=> adp + fdp
PFK_3	atp + s7p <=> adp + s17bp
PFL	coa + pyr <=> accoa + for

DSBAO2	mqn8 + dsbard_p <=> mql8 + dsbaox_p
EDA	2ddg6p <=> g3p + pyr
EDD	6pgc <=> 2ddg6p + h2o
ENO	2pg <=> h2o + pep
F6PA	f6p <=> dha + g3p
FBA	fdp <=> dhap + g3p
FBA3	s17bp <=> dhap + e4p
FBP	fdp + h2o <=> f6p + h + pi
FDH4pp	2 h + q8 + for_p <=> q8h2 + co2_p + h_p
FDH5pp	2 h + mqn8 + for_p <=> mql8 + co2_p + h_p
FHL	for + h <=> co2 + h2
FLDR2	2 flxso + h + nadph <=> 2 flxr + nadp
FORt2pp	for_p + h_p <=> for + h
FORtex	for_e <=> for_p
FORtppi	for <=> for_p
FRD2	fum + mql8 <=> mqn8 + succ
FRD3	2dmmql8 + fum <=> 2dmmq8 + succ
FUM	fum + h2o <=> mal-L
G1PPpp	g1p_p + h2o_p <=> glc-D_p + h_p + pi_p
G3PD2	glyc3p + nadp <=> dhap + h + nadph
G3PD5	glyc3p + q8 <=> dhap + q8h2
G3PD6	glyc3p + mqn8 <=> dhap + mql8
G3PD7	2dmmq8 + glyc3p <=> 2dmmql8 + dhap
G6PDH2r	g6p + nadp <=> 6pgl + h + nadph
G6PP	g6p + h2o <=> glc-D + h + pi
GAPD	g3p + nad + pi <=> 13dpg + nadh
GLBRAN2	glycogen <=> bglycogen
GLCDpp	q8 + glc-D_p + h2o_p <=> q8h2 + glcn_p + h_p
GLCP	glycogen + h + pi <=> g1p + glycogenn1
GLCP2	bglycogen + pi <=> bglycogenn1 + g1p
PGI	g6p <=> f6p
PGK	3pg + atp + h <=> 13dpg + adp
PGL	6pgl + h2o <=> 6pgc + h
PGM	2pg <=> 3pg
PGMT	g1p <=> g6p
PIt2rpp	h_p + pi_p <=> h + pi
PIuabcpp	atp + h2o + pi_p <=> adp + h + 2 pi
POR5	coa + 2 flxso + h + pyr <=> accoa + co2 + 2 flxr
POX	h2o + pyr + q8 <=> ac + co2 + q8h2
PPA	h2o + ppi <=> 2 h + 2 pi
PPA2	h2o + pppi <=> pi + ppi
PPC	co2 + h2o + pep <=> 2 h + oaa + pi
PPCK	atp + h + oaa <=> adp + co2 + pep
PPK2r	atp + ppi <=> adp + h + pppi
PPKr	atp + h + pi <=> adp + ppi
PPS	atp + h2o + pyr <=> amp + h + pep + pi
PTAr	accoa + h + pi <=> actp + coa
PYK	adp + pep <=> atp + pyr
QMO2	2 o2 + q8h2 <=> 2 o2s + q8
QMO3	mql8 + 2 o2 <=> mqn8 + 2 o2s
RPE	ru5p-D <=> xu5p-D
RPI	r5p <=> ru5p-D
SELR	mql8 + sel <=> h2o + mqn8 + slnt
SPODM	2 o2s <=> h2o2 + o2
SUCDi	q8 + succ <=> fum + q8h2
SUCOAS	atp + coa + succ <=> adp + pi + succoa
TALA	g3p + s7p <=> e4p + f6p
THD2pp	nadh + nadp + 2 h_p <=> 2 h + nad + nadph
THIORDXi	h2o2 + trdrd <=> 2 h2o + trdox
TKT1	r5p + xu5p-D <=> g3p + s7p

GLCS1	adpglc + glycogenn1 <=> adp + glycogen	TKT2	e4p + xu5p-D <=> f6p + g3p
GLCabcpp	atp + h2o + glc-D_p <=> adp + glc-D + h + pi	TMAOR1	h + mql8 + tmao <=> h2o + mqn8 + tma
GLCptspp	pep + glc-D_p <=> g6p + pyr	TMAOR1pp	mql8 + h_p + tmao_p <=> mqn8 + h2o_p + tma_p
GLCt2pp	glc-D_p + h_p <=> glc-D + h	TMAOR2	2dmmql8 + h + tmao <=> 2dmmq8 + h2o + tma
GLDBRAN2	bglycogen <=> glycogen	TMAOR2pp	2dmmql8 + h_p + tmao_p <=> 2dmmq8 + h2o_p + tma_p
GLGC	atp + g1p <=> adpglc + ppi	TPI	dhap <=> g3p
GLXCL	2 glx + h <=> 2h3oppan + co2	TRDR	h + nadph + trdox <=> nadp + trdrd
GLYCK	atp + glyc-R <=> 3pg + adp	TRSARr	2h3oppan + h + nadh <=> glyc-R + nad

Table A.4: Comparison of Flux Variability of ETC reactions in rGEM

Reactions	Reaction Formulas	rGEM min	rGEM max	GEM min	GEM max
ASPO3	asp-L + q8 <=> h + iasp + q8h2	0.000	0.002	0.000	0.002
ASPO4	asp-L + mqn8 <=> h + iasp + mql8	0.000	0.002	0.000	0.002
ATPSynthase	adp + pi + 4 h_p <=> atp + 3 h + h2o	-40.000	60.000	-40.000	60.000
CYTbd2	2 h + mql8 + 0.5 o2 <=> h2o + mqn8 + 2 h_p	0.000	60.000	0.000	60.000
CYTbd1	2 h + 0.5 o2 + q8h2 <=> h2o + q8 + 2 h_p	0.000	60.000	0.000	60.000
CYTbo3	4 h + 0.5 o2 + q8h2 <=> h2o + q8 + 4 h_p	0.000	60.000	0.000	60.000
DHORD2	dhor-S + q8 <=> orot + q8h2	0.000	0.319	0.000	16.116
DHORD5	dhor-S + mqn8 <=> mql8 + orot	0.000	0.319	0.000	16.116
FDH4pp	2 h + q8 + for_p <=> q8h2 + co2_p + h_p	0.000	60.000	0.000	60.000
FDH5pp	2 h + mqn8 + for_p <=> mql8 + co2_p + h_p	0.000	60.000	0.000	60.000
FRD3	2dmmql8 + fum <=> 2dmmq8 + succ	0.000	60.000	0.000	60.000
G3PD5	glyc3p + q8 <=> dhap + q8h2	0.000	60.000	0.000	60.000
G3PD6	glyc3p + mqn8 <=> dhap + mql8	0.000	60.000	0.000	60.000
G3PD7	2dmmq8 + glyc3p <=> 2dmmql8 + dhap	0.000	60.000	0.000	60.000
GLCDpp	q8 + glc-D_p + h2o_p <=> q8h2 + glcn_p + h_p	0.000	10.000	0.000	10.000
GLYCTO2	glyclt + q8 <=> glx + q8h2	0.000	60.000	0.000	60.000
GLYCTO3	glyclt + mqn8 <=> glx + mql8	0.000	60.000	0.000	60.000
GLYCTO4	2dmmq8 + glyclt <=> 2dmmql8 + glx	0.000	60.000	0.000	60.000
LDH_D2	lac-D + q8 <=> pyr + q8h2	0.000	60.000	0.000	60.000
MDH2	mal-L + q8 <=> oaa + q8h2	0.000	60.000	0.000	60.000
MDH3	mal-L + mqn8 <=> mql8 + oaa	0.000	60.000	0.000	60.000
NADH10	h + mqn8 + nadh <=> mql8 + nad	0.000	60.000	0.000	60.000
NADH16pp	4 h + nadh + q8 <=> nad + q8h2 + 3 h_p	0.000	60.000	0.000	60.000
NADH17pp	4 h + mqn8 + nadh <=> mql8 + nad + 3 h_p	0.000	60.000	0.000	60.000
NADH18pp	2dmmq8 + 4 h + nadh <=> 2dmmql8 + nad + 3 h_p	0.000	60.000	0.000	60.000
NADH5	h + nadh + q8 <=> nad + q8h2	0.000	60.000	0.000	60.000
NADH9	2dmmq8 + h + nadh <=> 2dmmql8 + nad	0.000	60.000	0.000	60.000
NADPHQR2	h + nadph + q8 <=> nadp + q8h2	0.000	60.000	0.000	60.000
NADPHQR3	h + mqn8 + nadph <=> mql8 + nadp	0.000	60.000	0.000	60.000
NADPHQR4	2dmmq8 + h + nadph <=> 2dmmql8 + nadp	0.000	60.000	0.000	60.000
NADTRHD	nad + nadph <=> nadh + nadp	0.000	60.000	0.000	60.000
POX	h2o + pyr + q8 <=> ac + co2 + q8h2	0.000	60.000	0.000	60.000
PPK2r	atp + ppi <=> adp + h + pppi	0.000	60.000	-60.000	60.000
PPKr	atp + h + pi <=> adp + ppi	-60.000	60.000	-60.000	60.000
SUCDi	q8 + succ <=> fum + q8h2	0.000	60.000	0.000	60.000
THD2	nadh + nadp + 2 h_p <=> 2 h + nad + nadph	0.000	60.000	0.000	60.000
TRDR	h + nadph + trdox <=> nadp + trdrd	0.000	58.529	0.000	60.000

Table A.5: Statistics of Subnetworks and Lumped Reactions of *E. coli* rGEM riJO1366 D=1

Biomass Building Blocks	Size of Subnetwork	# of Alternative Subnetwork	# of Unique Lumped Reactions
10-Formyltetrahydrofolate	42	4	4
2-Demethylmenaquinol 8	28	6	1
[2Fe-2S] iron-sulfur cluster	15	6	6
[4Fe-4S] iron-sulfur cluster	17	8	8
5-Methyltetrahydrofolate	42	6	2
Acetyl-CoA	43	6	2
Adenosylcobalamin	25	4	4
L-Alanine	1	1	1
S-Adenosyl-L-methionine	38	6	2
L-Arginine	9	1	1
L-Asparagine	1	1	1
bis-molybdopterin guanine dinucleotide	41	4	4
Biotin	64	24	8
Calcium	5	1	1
chorismate	7	1	1
Coenzyme A	43	6	2
CTP	10	1	1
L-Cysteine	11	2	2
dATP	24	4	4
dCTP	11	1	1
dGTP	25	6	6
dTTP	17	2	2
Enterochelin	16	2	1
Flavin adenine dinucleotide oxidized	34	4	4
L-Glutamine	1	1	1
Glycine	6	1	1
Reduced glutathione	17	2	2
GTP	24	4	4
Heme O	22	3	1
L-Histidine	17	2	2
L-Isoleucine	10	1	1
L-Leucine	9	1	1
lipoate (protein bound)	36	8	8
L-Lysine	9	1	1
Malonyl-CoA	44	6	2
L-Methionine	20	2	2
5,10-Methylenetetrahydrofolate	41	6	2

molybdopterin cytosine dinucleotide	47	8	8
molybdopterin guanine dinucleotide	39	4	4
Menaquinol 8	41	6	1
Nicotinamide adenine dinucleotide	27	4	4
Nicotinamide adenine dinucleotide - reduced	27	4	4
Nicotinamide adenine dinucleotide phosphate	27	4	4
Nicotinamide adenine dinucleotide phosphate - reduced	27	4	4
phosphatidylethanolamine (dihexadecanoyl, n-C16:0)	41	130	10
phosphatidylethanolamine (dihexadec-9enoyl, n-C16:1)	41	66	9
phosphatidylethanolamine (dioctadec-11-enoyl, n-C18:1)	45	130	10
Phosphatidylglycerol (dihexadecanoyl, n-C16:0)	38	130	10
Phosphatidylglycerol (dihexadec-9-enoyl, n-C16:1)	38	66	9
Phosphatidylglycerol (dioctadec-11-enoyl, n-C18:1)	42	64	10
L-Phenylalanine	10	1	1
Protoheme	9	1	1
L-Proline	4	1	1
Putrescine	6	1	1
Pyridoxal 5-phosphate	6	2	2
Ubiquinol-8	41	24	4
Riboflavin	30	4	4
L-Serine	3	1	1
Siroheme	21	1	1
Spermidine	33	2	2
Succinyl-CoA	43	6	2
5,6,7,8-Tetrahydrofolate	42	4	4
Thiamine diphosphate	42	2	2
L-Threonine	5	1	1
L-Tryptophan	15	1	1
L-Tyrosine	10	1	1
Undecaprenyl diphosphate	13	3	1
UTP	8	1	1
L-Valine	4	1	1
core oligosaccharide lipid A	73	64	12
cardiolipin (tetrahexadecanoyl, n-C16:0)	39	130	10
cardiolipin (tetrahexadec-9-enoyl, n-C16:1)	39	59	9
cardiolipin (tetraoctadec-11-enoyl, n-C18:1)	43	8	6
two linked disacharide tripeptide murein units (uncross-linked, middle of chain)	31	1	1
two disacharide linked murein units, tripeptide crosslinked tetrapeptide (A2pm->D-ala) (middle of chain)	31	1	1
two linked disacharide tetrapeptide murein units (uncross-linked, middle of chain)	31	1	1
two disacharide linked murein units, tetrapeptide corsslinked tetrapeptide (A2pm->D-ala) (middle of chain)	31	1	1

three disacharide linked murein units (tetrapeptide cross-linked tetrapeptide (A2pm->D-ala) & tetrapeptide corsslinked tetrapeptide (A2pm->D-ala)) (middle of chain)	32	1	1
Growth Associated Maintenance	23	4	4

Table A.6: Statistics of Subnetworks and Lumped Reactions of *P. putida* rGEM D=1

Biomass Building Blocks	Size of Subnetwork	# of Alternative Subnetwork	# of Unique Lumped Reactions
Nicotinamide adenine dinucleotide phosphate - reduced	28	2	1
Nicotinamide adenine dinucleotide phosphate	28	2	1
L-Alanine	3	2	2
L-Glutamate	2	1	1
AMP	22	2	1
L-Asparagine	4	1	1
L-Aspartate	2	1	1
L-Glutamine	3	2	2
Acetyl-CoA	52	8	4
UTP	12	1	1
L-Arginine	11	3	3
Nicotinamide adenine dinucleotide	28	2	1
Nicotinamide adenine dinucleotide - reduced	28	2	1
L-Proline	4	2	2
Putrescine	7	1	1
L-Cysteine	8	1	1
L-Serine	5	1	1
phosphatidylethanolamine (dihexadec-9enoyl, n-C16:1)	42	2	2
cyclopropane phosphatidylethanolamine (dihexadec-9,10-cyclo-anoyl, n-C16:0 cyclo)	46	4	2
cyclopropane phosphatidylglycerol (dihexadec-9,10-cyclo-anoyl, n-C16:0 cyclo)	41	4	2
phosphatidylethanolamine (dioctadec-11-enoyl, n-C18:1)	46	4	2
cyclopropane phosphatidylethanolamine (dioctadec-11,12-cyclo-anoyl, n-C18:0 cyclo)	50	4	2
cyclopropane phosphatidylglycerol (dioctadec-11,12-cyclo-anoyl, n-C18:0 cyclo)	45	4	2
GTP	23	2	1
Glycine	7	4	2
CTP	13	1	1
phosphatidylethanolamine (didodecanoyl, n-C12:0)	33	2	1
phosphatidylethanolamine (dihexadecanoyl, n-C16:0)	42	4	2
phosphatidylethanolamine (dioctadecanoyl, n-C18:0)	46	4	2
Phosphatidylglycerol (didodecanoyl, n-C12:0)	28	2	1
Phosphatidylglycerol (dihexadecanoyl, n-C16:0)	37	4	2

Phosphatidylglycerol (dioctadecanoyl, n-C18:0)	41	4	2
L-Methionine	18	8	6
L-Threonine	7	1	1
UDPglucose	12	1	1
L-Histidine	22	5	5
L-Phenylalanine	12	1	1
L-Tyrosine	12	1	1
L-Lysine	12	3	3
5-Methyltetrahydrofolate	43	10	5
Peptidoglycan subunit	32	17	17
L-Tryptophan	19	2	2
Heme O	25	1	1
Siroheme	14	1	1
dGTP	25	2	1
dATP	25	2	1
dCTP	15	1	1
dTTP	24	4	4
L-Isoleucine	11	1	1
L-Leucine	10	2	1
L-Valine	6	1	1
Cardiolipin (tetrahexadecanoyl, n-C12:0)	30	2	1
Cardiolipin (tetrahexadecanoyl, n-C16:0)	39	4	2
Cardiolipin (tetrahexadec-9-enoyl, n-C16:1)	39	4	2
Cardiolipin (tetraoctadecanoyl, n-C18:0)	43	4	2
Cardiolipin (tetraoctadec-11-enoyl, n-C18:1)	43	4	2
Growth Associated Maintanence	22	2	1

Table A.7: Statistics of Subnetworks and Lumped Reactions of *S. cerevisiae* rGEM D=1

Biomass Building Blocks	Size of Subnetwork	# of Alternative Subnetwork	# of Unique Lumped Reactions
1_3_beta_D_Glucan	3	1	1
L_Alanine	2	2	2
AMP	23	26	26
L_Arginine	12	2	2
L_Asparagine	4	3	3
L_Aspartate	2	2	2
CMP	14	6	6
L_Cysteine	12	6	6
dAMP	26	26	26
dCMP	14	4	4
dGMP	26	30	30
dTMP	19	12	12
Ergosterol	33	2	2
L_Glutamine	2	1	1
L_Glutamate	2	1	1
Glycine	2	2	2
Glycogen	3	1	1
GMP	22	13	13
L_Histidine	24	28	28
L_Isoleucine	12	12	12
L_Leucine	11	2	2
L_Lysine	10	2	2
Mannan	6	2	2
L_Methionine	19	24	24
Phosphatidate__yeast_specific	13	2	2
Phosphatidylcholine__yeast_specific	31	6	6
Phosphatidylethanolamine__yeast_specific	22	6	6
L_Phenylalanine	11	1	1
L_Proline	5	2	2
Phosphatidylserine__yeast_specific	21	6	6
Phosphatidyl_1D_myo_inositol__yeast_specific	19	2	2
Riboflavin	26	14	14
L_Serine	4	3	3
L_Threonine	7	4	4
Trehalose	4	1	1
triglyceride__yeast_specific	15	2	2
L_Tryptophan	18	5	5
L_Tyrosine	11	2	2

UMP	9	3	3
L_Valine	6	1	1
Zymosterol	22	2	2

Table A.8: Statistics of Subnetworks and Lumped Reactions of *Homo sapiens* rGEM D=1

Biomass Building Blocks	Size of Subnetwork	# of Alternative Subnetwork	# of Unique Lumped Reactions
L-Glutamate	2	2	2
L-Aspartate	2	4	4
GTP	21	8	8
L-Asparagine	4	4	4
L-Alanine	2	2	2
L-Glutamine	2	2	2
Glycine	3	2	2
L-Serine	4	2	2
L-Arginine	8	32	16
CTP	10	8	8
Cholesterol	23	64	3
UTP	9	4	4
dGTP	23	16	16
dCTP	12	8	8
dATP	23	8	8
dTTP	20	48	48
L-Proline	4	6	5

BIBLIOGRAPHY

1. DeBerardinis RJ, Thompson CB: **Cellular Metabolism and Disease: What Do Metabolic Outliers Teach Us?** *Cell* 2012, **148**:1132-1144.
2. Gill SR, Pop M, DeBoy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE: **Metagenomic analysis of the human distal gut microbiome**. *Science* 2006, **312**:1355-1359.
3. Guinane CM, Cotter PD: **Role of the gut microbiota in health and chronic gastrointestinal disease: understanding a hidden metabolic organ**. *Therapeutic Advances in Gastroenterology* 2013, **6**:295-308.
4. Frazzetto G: **White biotechnology - The application of biotechnology to industrial production holds many promises for sustainable development, but many products still have to pass the test of economic viability**. *Embo Reports* 2003, **4**:835-837.
5. Heavner BD, Smallbone K, Price ND, Walker LP: **Version 6 of the consensus yeast metabolic network refines biochemical coverage and improves model performance**. *Database : the journal of biological databases and curation* 2013, **2013**:bat059.
6. Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, Palsson BO: **A comprehensive genome-scale reconstruction of Escherichia coli metabolism--2011**. *Molecular systems biology* 2011, **7**:535.
7. Zhao J, Yu H, Luo JH, Cao ZW, Li YX: **Hierarchical modularity of nested bow-ties in metabolic networks**. *Bmc Bioinformatics* 2006, **7**:386.
8. Francke C, Siezen RJ, Teusink B: **Reconstructing the metabolic network of a bacterium from its genome**. *Trends in microbiology* 2005, **13**:550-558.

9. Price ND, Papin JA, Schilling CH, Palsson BO: **Genome-scale microbial in silico models: the constraints-based approach**. *Trends in biotechnology* 2003, **21**:162-169.
10. Feist AM, Palsson BO: **The biomass objective function**. *Current Opinion in Microbiology* 2010, **13**:344-349.
11. Bonarius HPJ, Schmid G, Tramper J: **Flux analysis of underdetermined metabolic networks: The quest for the missing constraints**. *Trends in biotechnology* 1997, **15**:308-314.
12. Henry CS, Jankowski MD, Broadbelt LJ, Hatzimanikatis V: **Genome-scale thermodynamic analysis of Escherichia coli metabolism**. *Biophysical Journal* 2006, **90**:1453-1461.
13. Soh KC, Hatzimanikatis V: **Constraining the flux space using thermodynamics and integration of metabolomics data**. *Methods in molecular biology (Clifton, N.J.)* 2014, **1191**:49-63.
14. Henry CS, Broadbelt LJ, Hatzimanikatis V: **Thermodynamics-based metabolic flux analysis**. *Biophysical journal* 2007, **92**:1792-1805.
15. Ataman M, Hatzimanikatis V: **Heading in the right direction: thermodynamics-based network analysis and pathway engineering**. *Current opinion in biotechnology* 2015, **36**:176-182.
16. Miskovic L, Hatzimanikatis V: **Production of biofuels and biochemicals: in need of an ORACLE**. *Trends in biotechnology* 2010, **28**:391-397.
17. Heijnen JJ, Van Dijken JP: **In search of a thermodynamic description of biomass yields for the chemotrophic growth of microorganisms**. *Biotechnology and Bioengineering* 1992, **39**:833-858.
18. Famili I, Forster J, Nielsen J, Palsson BO: **Saccharomyces cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**:13134-13139.
19. Ma HW, Zhao XM, Yuan YJ, Zeng AP: **Decomposition of metabolic network into functional modules based on the global connectivity structure of reaction graph**. *Bioinformatics* 2004, **20**:1870-1876.

20. Selkov E, Maltsev N, Olsen GJ, Overbeek R, Whitman WB: **A reconstruction of the metabolism of Methanococcus jannaschii from sequence data (Reprinted from Gene-Combis, vol 197, pg GC11-GC26, 1997)**. *Gene* 1997, **197**:Gc11-Gc26.

21. Ingraham JL, Maaløe O, Neidhardt FC: *Growth of the bacterial cell*. Sinauer Associates; 1983.

22. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M: **KEGG: Kyoto Encyclopedia of Genes and Genomes**. *Nucleic acids research* 1999, **27**:29-34.

23. Keseler IM, Collado-Vides J, Santos-Zavaleta A, Peralta-Gil M, Gama-Castro S, Muniz-Rascado L, Bonavides-Martinez C, Paley S, Krummenacker M, Altman T, et al.: **EcoCyc: a comprehensive database of Escherichia coli biology**. *Nucleic acids research* 2011, **39**:D583-590.

24. Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL: **High-throughput generation, optimization and analysis of genome-scale metabolic models**. *Nature Biotechnology* 2010, **28**:977-U922.

25. Nielsen J, Jewett MC: **Impact of systems biology on metabolic engineering of Saccharomyces cerevisiae**. *FEMS yeast research* 2008, **8**:122-131.

26. Kell DB, Oliver SG: **Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era**. *BioEssays : news and reviews in molecular, cellular and developmental biology* 2004, **26**:99-105.

27. Soh KC, Miskovic L, Hatzimanikatis V: **From network models to network responses: integration of thermodynamic and kinetic properties of yeast genome-scale metabolic networks**. *FEMS yeast research* 2012, **12**:129-143.

28. Zamboni N, Fendt SM, Ruhl M, Sauer U: **(13)C-based metabolic flux analysis**. *Nature protocols* 2009, **4**:878-892.

29. Hatzimanikatis V, Li C, Ionita JA, Henry CS, Jankowski MD, Broadbelt LJ: **Exploring the diversity of complex metabolic networks**. *Bioinformatics* 2005, **21**:1603-1609.

30. Tymoshenko S, Oppenheim RD, Agren R, Nielsen J, Soldati-Favre D, Hatzimanikatis V: **Metabolic Needs and Capabilities of Toxoplasma gondii through**

Combined Computational and Experimental Analysis. *Plos Computational Biology* 2015, **11**:e1004261.

31. Bordbar A, Monk JM, King ZA, Palsson BO: **Constraint-based models predict metabolic and associated cellular functions**. *Nature reviews. Genetics* 2014, **15**:107-120.
32. Xu C, Liu L, Zhang Z, Jin D, Qiu J, Chen M: **Genome-scale metabolic model in guiding metabolic engineering of microbial improvement**. *Applied microbiology and biotechnology* 2013, **97**:519-539.
33. Harcombe WR, Delaney NF, Leiby N, Klitgord N, Marx CJ: **The ability of flux balance analysis to predict evolution of central metabolism scales with the initial distance to the optimum**. *PLoS computational biology* 2013, **9**:e1003091.
34. Soh KC, Hatzimanikatis V: **Network thermodynamics in the post-genomic era**. *Current Opinion in Microbiology* 2010, **13**:350-357.
35. Beard DA, Liang SD, Qian H: **Energy balance for analysis of complex metabolic networks**. *Biophysical journal* 2002, **83**:79-86.
36. Kummel A, Panke S, Heinemann M: **Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data**. *Molecular systems biology* 2006, **2**:2006 0034.
37. Zamboni N, Kummel A, Heinemann M: **anNET: a tool for network-embedded thermodynamic analysis of quantitative metabolome data**. *Bmc Bioinformatics* 2008, **9**.
38. Bennett BD, Kimball EH, Gao M, Osterhout R, Van Dien SJ, Rabinowitz JD: **Absolute metabolite concentrations and implied enzyme active site occupancy in Escherichia coli**. *Nature chemical biology* 2009, **5**:593-599.
39. Dreyfuss JM, Zucker JD, Hood HM, Ocasio LR, Sachs MS, Galagan JE: **Reconstruction and validation of a genome-scale metabolic model for the filamentous fungus Neurospora crassa using FARM**. *Plos Computational Biology* 2013, **9**:e1003126.
40. Jankowski MD, Henry CS, Broadbelt LJ, Hatzimanikatis V: **Group contribution method for thermodynamic analysis of complex metabolic networks**. *Biophysical journal* 2008, **95**:1487-1499.

41. Mavrovouniotis: **Estimation of standard Gibbs energy changes of biotransformations**. *The Journal of biological chemistry* 1991, **266**:14440-14445.

42. Mavrovouniotis ML: **Group contributions for estimating standard gibbs energies of formation of biochemical compounds in aqueous solution**. *Biotechnol. Bioeng.* 1990, **36**:1070-1082.

43. Pitkanen E, Jouhten P, Hou J, Syed MF, Blomberg P, Kludas J, Oja M, Holm L, Penttila M, Rousu J, et al.: **Comparative Genome-Scale Reconstruction of Gapless Metabolic Networks for Present and Ancestral Species**. *Plos Computational Biology* 2014, **10**.

44. Flamholz A, Noor E, Bar-Even A, Milo R: **eQuilibrator-the biochemical thermodynamics calculator**. *Nucleic acids research* 2012, **40**:D770-D775.

45. Karlstaedt A, Fliegner D, Kararigas G, Ruderisch HS, Regitz-Zagrosek V, Holzhuetter HG: **CardioNet: A human metabolic network suited for the study of cardiomyocyte metabolism**. *European Journal of Heart Failure* 2013, **12**:S42-S42.

46. Henry CS, Zinner JF, Cohoon MP, Stevens RL: **iBsu1103: a new genome-scale metabolic model of Bacillus subtilis based on SEED annotations**. *Genome Biology* 2009, **10**.

47. Gille C, Bolling C, Hoppe A, Bulik S, Hoffmann S, Hubner K, Karlstadt A, Ganeshan R, Konig M, Rother K, et al.: **HepatoNet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology**. *Molecular Systems Biology* 2010, **6**.

48. Thiele I, Hyduke DR, Steeb B, Fankam G, Allen DK, Bazzani S, Charusanti P, Chen FC, Fleming RMT, Hsiung CA, et al.: **A community effort towards a knowledge-base and mathematical model of the human pathogen Salmonella Typhimurium LT2**. *Bmc Systems Biology* 2011, **5**.

49. Tanaka K, Henry CS, Zinner JF, Jolivet E, Cohoon MP, Xia F, Bidnenko V, Ehrlich SD, Stevens RL, Noirot P: **Building the repertoire of dispensable chromosome regions in Bacillus subtilis entails major refinement of cognate large-scale metabolic model**. *Nucleic acids research* 2013, **41**:687-699.

50. Imam S, Yilmaz S, Sohmen U, Gorzalski AS, Reed JL, Noguera DR, Donohue TJ: **iRsp1095: a genome-scale reconstruction of the Rhodobacter sphaeroides metabolic network**. *Bmc Systems Biology* 2011, **5**:116.

51. McAnulty MJ, Yen JY, Freedman BG, Senger RS: **Genome-scale modeling using flux ratio constraints to enable metabolic engineering of clostridial metabolism in silico**. *Bmc Systems Biology* 2012, **6**.

52. Martinez VS, Quek LE, Nielsen LK: **Network thermodynamic curation of human and yeast genome-scale metabolic models**. *Biophysical journal* 2014, **107**:493-503.

53. Schultz A, Qutub AA: **Predicting internal cell fluxes at sub-optimal growth**. *Bmc Systems Biology* 2015, **9**.

54. De Martino D, Figliuzzi M, De Martino A, Marinari E: **A Scalable Algorithm to Explore the Gibbs Energy Landscape of Genome-Scale Metabolic Networks**. *Plos Computational Biology* 2012, **8**.

55. McCloskey D, Gangoiti JA, King ZA, Naviaux RK, Barshop BA, Palsson BO, Feist AM: **A Model-Driven Quantitative Metabolomics Analysis of Aerobic and Anaerobic Metabolism in E. coli K-12 MG1655 That Is Biochemically and Thermodynamically Consistent**. *Biotechnology and Bioengineering* 2014, **111**:803-815.

56. Orman MA, Androulakis IP, Berthiaume F, Ierapetritou MG: **Metabolic network analysis of perfused livers under fed and fasted states: Incorporating thermodynamic and futile-cycle-associated regulatory constraints**. *Journal of Theoretical Biology* 2012, **293**:101-110.

57. Campodonico MA, Andrews BA, Asenjo JA, Palsson BO, Feist AM: **Generation of an atlas for commodity chemical production in Escherichia coli and a novel pathway prediction algorithm, GEM-Path**. *Metabolic engineering* 2014, **25**:140-158.

58. Araki M, Cox RS, Makiguchi H, Ogawa T, Taniguchi T, Miyaoku K, Nakatsui M, Hara KY, Kondo A: **M-path: a compass for navigating potential metabolic pathways**. *Bioinformatics* 2015, **31**:905-911.

59. Carbonell P, Parutto P, Herisson J, Pandit SB, Faulon JL: **XTMS: pathway design in an eXTended metabolic space**. *Nucleic acids research* 2014, **42**:W389-W394.
60. McClymont K, Soyer OS: **Metabolic tinker: an online tool for guiding the design of synthetic metabolic pathways**. *Nucleic acids research* 2013, **41**:e113.
61. Pertusi DA, Stine AE, Broadbelt LJ, Tyo KEJ: **Efficient searching and annotation of metabolic networks using chemical similarity**. *Bioinformatics* 2015, **31**:1016-1024.
62. Hadadi N, Soh KC, Seijo M, Zisaki A, Guan XL, Wenk MR, Hatzimanikatis V: **A computational framework for integration of lipidomics data into metabolic pathways**. *Metabolic engineering* 2014, **23**:1-8.
63. Henry CS, Broadbelt LJ, Hatzimanikatis V: **Discovery and Analysis of Novel Metabolic Pathways for the Biosynthesis of Industrial Chemicals: 3-Hydroxypropanoate**. *Biotechnology and Bioengineering* 2010, **106**:462-473.
64. Zhu Y, Song JN, Xu ZX, Sun JB, Zhang YP, Li Y, Ma YH: **Development of thermodynamic optimum searching (TOS) to improve the prediction accuracy of flux balance analysis**. *Biotechnology and Bioengineering* 2013, **110**:914-923.
65. Muller AC, Bockmayr A: **Fast thermodynamically constrained flux variability analysis**. *Bioinformatics* 2013, **29**:903-909.
66. Zanghellini J, Ruckerbauer DE, Hanscho M, Jungreuthmayer C: **Elementary flux modes in a nutshell: Properties, calculation and applications**. *Biotechnology journal* 2013, **8**:1009-U1061.
67. Gerstl MP, Jungreuthmayer C, Zanghellini J: **tEFMA: computing thermodynamically feasible elementary flux modes in metabolic networks**. *Bioinformatics* 2015.
68. Jol SJ, Kummel A, Terzer M, Stelling J, Heinemann M: **System-Level Insights into Yeast Metabolism by Thermodynamic Analysis of Elementary Flux Modes**. *Plos Computational Biology* 2012, **8**.
69. Muller AC, Bockmayr A: **Flux modules in metabolic networks**. *Journal of Mathematical Biology* 2014, **69**:1151-1179.

70. Jorda J, Suarez C, Carnicer M, ten Pierick A, Heijnen JJ, van Gulik W, Ferrer P, Albiol J, Wahl A: **Glucose-methanol co-utilization in Pichia pastoris studied by metabolomics and instationary (1)(3)C flux analysis**. *Bmc Systems Biology* 2013, **7**:17.

71. Tepper N, Noor E, Amador-Noguez D, Haraldsdottir HS, Milo R, Rabinowitz J, Liebermeister W, Shlomi T: **Steady-state metabolite concentrations reflect a balance between maximizing enzyme efficiency and minimizing total metabolite load**. *PloS one* 2013, **8**:e75370.

72. Birkenmeier M, Mack M, Roder T: **A coupled thermodynamic and metabolic control analysis methodology and its evaluation on glycerol biosynthesis in Saccharomyces cerevisiae**. *Biotechnology letters* 2014.

73. Chakrabarti A, Miskovic L, Soh KC, Hatzimanikatis V: **Towards kinetic modeling of genome-scale metabolic networks without sacrificing stoichiometric, thermodynamic and physiological constraints**. *Biotechnology journal* 2013, **8**:1043-U1105.

74. Noor E, Bar-Even A, Flamholz A, Reznik E, Liebermeister W, Milo R: **Pathway Thermodynamics Highlights Kinetic Obstacles in Central Metabolism**. *Plos Computational Biology* 2014, **10**.

75. Flamholz A, Noor E, Bar-Even A, Liebermeister W, Milo R: **Glycolytic strategy as a tradeoff between energy yield and protein cost**. *Proceedings of the National Academy of Sciences of the United States of America* 2013, **110**:10039-10044.

76. Simons M, Saha R, Amiour N, Kumar A, Guillard L, Clement G, Miquel M, Li ZN, Mouille G, Lea PJ, et al.: **Assessing the Metabolic Impact of Nitrogen Availability Using a Compartmentalized Maize Leaf Genome-Scale Model**. *Plant Physiology* 2014, **166**:1659-1674.

77. Seaver SMD, Gerdes S, Frelin O, Lerma-Ortiz C, Bradbury LMT, Zallot R, Hasnain G, Niehaus TD, El Yacoubi B, Pasternak S, et al.: **High-throughput comparison, functional annotation, and metabolic modeling of plant genomes using the PlantSEED resource**. *Proceedings of the National Academy of Sciences of the United States of America* 2014, **111**:9645-9650.

78. Jol SJ, Kummel A, Hatzimanikatis V, Beard DA, Heinemann M: **Thermodynamic Calculations for Biochemical Transport and Reaction Processes in Metabolic Networks**. *Biophysical journal* 2010, **99**:3139-3144.
79. Paget CM, Schwartz JM, Delneri D: **Environmental systems biology of cold-tolerant phenotype in Saccharomyces species adapted to grow at different temperatures**. *Molecular Ecology* 2014, **23**:5241-5257.
80. Hadadi N, Ataman M, Hatzimanikatis V, Panayiotou C: **Molecular thermodynamics of metabolism: quantum thermochemical calculations for key metabolites**. *Physical chemistry chemical physics : PCCP* 2015, **17**:10438-10453.
81. Jinich A, Rappoport D, Dunn I, Sanchez-Lengeling B, Olivares-Amaya R, Noor E, Even AB, Aspuru-Guzik A: **Quantum chemical approach to estimating the thermodynamics of metabolic reactions**. *Scientific Reports* 2014, **4**:7022.
82. Nicholls DG, Ferguson SJ: *Bioenergetics 3*. Academic Press; 2002.
83. Harris DA: *Bioenergetics at a glance*. Blackwell Science; 1995.
84. Russell JB, Cook GM: **Energetics of Bacterial-Growth - Balance of Anabolic and Catabolic Reactions**. *Microbiological Reviews* 1995, **59**:48-62.
85. Rottenberg H: **Non-equilibrium thermodynamics of energy conversion in bioenergetics**. *Biochimica et biophysica acta* 1979, **549**:225-253.
86. Aledo JC, del Valle AE: **Glycolysis in wonderland: The importance of energy dissipation in metabolic pathways**. *Journal of Chemical Education* 2002, **79**:1336-1339.
87. Anraku Y: **Bacterial Electron-Transport Chains**. *Annual Review of Biochemistry* 1988, **57**:101-132.
88. Haddock BA, Jones CW: **Bacterial Respiration**. *Bacteriological Reviews* 1977, **41**:47-99.
89. Hanegraaf PPF, Stouthamer AH, Kooijman SALM: **A mathematical model for yeast respiro-fermentative physiology**. *Yeast* 2000, **16**:423-437.
90. von Stockar U, Liu JS: **Does microbial life always feed on negative entropy? Thermodynamic analysis of microbial growth**. *Biochimica Et Biophysica Acta-Bioenergetics* 1999, **1412**:191-211.

91. Noguchi Y, Nakai Y, Shimba N, Toyosaki H, Kawahara Y, Sugimoto S, Suzuki E: **The energetic conversion competence of Escherichia coli during aerobic respiration studied by P-31 NMR using a circulating fermentation system**. *Journal of Biochemistry* 2004, **136**:509-515.

92. Hinkle PC, Yu ML: **The phosphorus/oxygen ratio of mitochondrial oxidative phosphorylation**. *The Journal of biological chemistry* 1979, **254**:2450-2455.

93. Hinkle PC: **P/O ratios of mitochondrial oxidative phosphorylation**. *Biochimica et biophysica acta* 2005, **1706**:1-11.

94. Smith AL, Hansen M: **Evidence for P/O ratios approaching 6 in mitochondrial oxidative phosphorylation**. *Biochemical and biophysical research communications* 1964, **15**:431-435.

95. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BO: **A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information**. *Molecular Systems Biology* 2007, **3**.

96. Taymaz-Nikerel H, Borujeni AE, Verheijen PJ, Heijnen JJ, van Gulik WM: **Genome-derived minimal metabolic models for Escherichia coli MG1655 with estimated in vivo respiratory ATP stoichiometry**. *Biotechnology and Bioengineering* 2010, **107**:369-381.

97. van Rijsewijk BRBH, Nanchen A, Nallet S, Kleijn RJ, Sauer U: **Large-scale C-13-flux analysis reveals distinct transcriptional control of respiratory and fermentative metabolism in Escherichia coli**. *Molecular Systems Biology* 2011, **7**.

98. Bekker M, de Vries S, Ter Beek A, Hellingwerf KJ, de Mattos MJT: **Respiration of Escherichia coli Can Be Fully Uncoupled via the Nonelectrogenic Terminal Cytochrome bd-II Oxidase**. *Journal of bacteriology* 2009, **191**:5510-5517.

99. Zhuang K, Vemuri GN, Mahadevan R: **Economics of membrane occupancy and respiro-fermentation**. *Molecular systems biology* 2011, **7**:500.

100. Schuetz R, Kuepfer L, Sauer U: **Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli**. *Molecular systems biology* 2007, **3**:119.

101. Luli GW, Strohl WR: **Comparison of growth, acetate production, and acetate inhibition of Escherichia coli strains in batch and fed-batch fermentations**. *Applied and Environmental Microbiology* 1990, **56**:1004-1011.

102. Farmer WR, Liao JC: **Reduction of aerobic acetate production by Escherichia coli**. *Applied and Environmental Microbiology* 1997, **63**:3205-3210.

103. Papoutsakis ET: **Equations and calculations for fermentations of butyric acid bacteria**. *Biotechnology and bioengineering* 1984, **26**:174-187.

104. Varma A, Boesch BW, Palsson BO: **Stoichiometric interpretation of Escherichia coli glucose catabolism under various oxygenation rates**. *Applied and Environmental Microbiology* 1993, **59**:2465-2473.

105. Vallino JJ, Stephanopoulos G: **Carbon Flux Distributions at the Pyruvate Branch Point in Corynebacterium-Glutamicum during Lysine Overproduction**. *Biotechnology progress* 1994, **10**:320-326.

106. Varma A, Palsson BO: **Metabolic capabilities of Escherichia coli: I. synthesis of biosynthetic precursors and cofactors**. *Journal of Theoretical Biology* 1993, **165**:477-502.

107. Varma A, Palsson BO: **Metabolic Capabilities of Escherichia-Coli .2. Optimal-Growth Patterns**. *Journal of Theoretical Biology* 1993, **165**:503-522.

108. Macnab RM: **Physiology of the Bacterial-Cell - a Molecular Approach - Neidhart,Fc, Ingraham,Jl, Schaechter,M**. *Nature* 1990, **348**:401-401.

109. Pramanik J, Keasling JD: **Stoichiometric model of Escherichia coli metabolism: Incorporation of growth-rate dependent biomass composition and mechanistic energy requirements**. *Biotechnology and Bioengineering* 1997, **56**:398-421.

110. Varma A, Palsson BO: **Stoichiometric Flux Balance Models Quantitatively Predict Growth and Metabolic by-Product Secretion in Wild-Type Escherichia-Coli W3110**. *Applied and Environmental Microbiology* 1994, **60**:3724-3731.

111. Pramanik J, Keasling JD: **Effect of Escherichia coli biomass composition on central metabolic fluxes predicted by a stoichiometric model**. *Biotechnology and Bioengineering* 1998, **60**:230-238.

112. Kim TY, Sohn SB, Kim YB, Kim WJ, Lee SY: **Recent advances in reconstruction and applications of genome-scale metabolic models**. *Current opinion in biotechnology* 2012, **23**:617-623.

113. Reed JL, Famili I, Thiele I, Palsson BO: **Towards multidimensional genome annotation**. *Nature Reviews Genetics* 2006, **7**:130-141.

114. Drell D: **The Department of Energy microbial cell project: A 180 degrees paradigm shift for biology**. *Omics : a journal of integrative biology* 2002, **6**:3-9.

115. Edwards JS, Palsson BO: **The Escherichia coli MG1655 in silico metabolic genotype: Its definition, characteristics, and capabilities**. *Proceedings of the National Academy of Sciences of the United States of America* 2000, **97**:5528-5533.

116. Durot M, Bourguignon PY, Schachter V: **Genome-scale models of bacterial metabolism: reconstruction and applications**. *Fems Microbiology Reviews* 2009, **33**:164-190.

117. Forster J, Famili I, Fu P, Palsson BO, Nielsen J: **Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network**. *Genome Research* 2003, **13**:244-253.

118. Schilling CH, Covert MW, Famili I, Church GM, Edwards JS, Palsson BO: **Genome-scale metabolic model of Helicobacter pylori 26695**. *Journal of bacteriology* 2002, **184**:4582-4593.

119. Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, Palsson BO: **A comprehensive genome-scale reconstruction of Escherichia coli metabolism-2011**. *Molecular Systems Biology* 2011, **7**.

120. McCloskey D, Palsson BO, Feist AM: **Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli**. *Molecular Systems Biology* 2013, **9**.

121. Sigurdsson MI, Jamshidi N, Steingrimsson E, Thiele I, Palsson BO: **A detailed genome-wide reconstruction of mouse metabolism based on human Recon 1**. *Bmc Systems Biology* 2010, **4**.

122. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BO: **Global reconstruction of the human metabolic network based on genomic**

and bibliomic data. *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**:1777-1782.

123. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crecy-Lagard V, Diaz N, Disz T, Edwards R, et al.: **The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes**. *Nucleic acids research* 2005, **33**:5691-5702.

124. Keseler IM, Mackie A, Peralta-Gil M, Santos-Zavaleta A, Gama-Castro S, Bonavides-Martinez C, Fulcher C, Huerta AM, Kothari A, Krummenacker M, et al.: **EcoCyc: fusing model organism databases with systems biology**. *Nucleic acids research* 2013, **41**:D605-612.

125. Lee S, Phalakornkule C, Domach MM, Grossmann IE: **Recursive MILP model for finding all the alternate optima in LP models for metabolic networks**. *Computers & Chemical Engineering* 2000, **24**:711-716.

126. Hatzimanikatis V, Floudas CA, Bailey JE: **Optimization of regulatory architectures in metabolic reaction networks**. *Biotechnology and Bioengineering* 1996, **52**:485-500.

127. Hatzimanikatis V, Floudas CA, Bailey JE: **Analysis and design of metabolic reaction networks via mixed-integer linear optimization**. *Aiche Journal* 1996, **42**:1277-1292.

128. Burgard AP, Maranas CD: **Probing the performance limits of the Escherichia coli metabolic network subject to gene additions or deletions**. *Biotechnology and Bioengineering* 2001, **74**:364-375.

129. de Figueiredo LF, Podhorski A, Rubio A, Kaleta C, Beasley JE, Schuster S, Planes FJ: **Computing the shortest elementary flux modes in genome-scale metabolic networks**. *Bioinformatics* 2009, **25**:3158-3165.

130. Borodina I, Nielsen J: **Advances in metabolic engineering of yeast Saccharomyces cerevisiae for production of chemicals**. *Biotechnology journal* 2014, **9**:609-620.

131. Christen S, Sauer U: **Intracellular characterization of aerobic glucose metabolism in seven yeast species by 13C flux analysis and metabolomics**. *FEMS yeast research* 2011, **11**:263-272.

132. Radrich K, Tsuruoka Y, Dobson P, Gevorgyan A, Swainston N, Baart G, Schwartz JM: **Integration of metabolic databases for the reconstruction of genome-scale metabolic networks**. *BMC systems biology* 2010, **4**:-.

133. DeJongh M, Formsma K, Boillot P, Gould J, Rycenga M, Best A: **Toward the automated generation of genome-scale metabolic networks in the SEED**. *BMC Bioinformatics* 2007, **8**:-.

134. Feist AM, Scholten JC, Palsson BO, Brockman FJ, Ideker T: **Modeling methanogenesis with a genome-scale metabolic reconstruction of Methanosarcina barkeri**. *Molecular systems biology* 2006, **2**:2006 0004.

135. Segre D, Vitkup D, Church GM: **Analysis of optimality in natural and perturbed metabolic networks**. *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99**:15112-15117.

136. Varma A, Palsson BO: **Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110**. *Applied and environmental microbiology* 1994, **60**:3724-3731.

137. Henry CS, Broadbelt LJ, Hatzimanikatis V: **Thermodynamics-based metabolic flux analysis**. In *Biophysical Journal*; 2007:1792-1805.

138. Alam MT, Medema MH, Takano E, Breitling R: **Comparative genome-scale metabolic modeling of actinomycetes: The topology of essential core metabolism**. *Febs Letters* 2011, **585**:2389-2394.

139. Baumler DJ, Peplinski RG, Reed JL, Glasner JD, Perna NT: **The evolution of metabolic networks of E. coli**. *Bmc Systems Biology* 2011, **5**.

140. Almaas E, Oltvai ZN, Barabasi AL: **The activity reaction core and plasticity of metabolic networks**. *PLoS computational biology* 2005, **1**:557-563.

141. Quek LE, Dietmair S, Hanscho M, Martinez VS, Borth N, Nielsen LK: **Reducing Recon 2 for steady-state flux analysis of HEK cell culture**. *Journal of Biotechnology* 2014, **184**:172-178.

142. Usuda Y, Nishio Y, Iwatani S, Van Dien SJ, Imaizumi A, Shimbo K, Kageyama N, Iwahata D, Miyano H, Matsui K: **Dynamic modeling of Escherichia coli**

metabolic and regulatory systems for amino-acid production. *Journal of Biotechnology* 2010, **147**:17-30.

143. Erdrich P, Steuer R, Klamt S: **An algorithm for the reduction of genome-scale metabolic network models to meaningful core models**. *Bmc Systems Biology* 2015, **9**:48.

144. Planes FJ, Beasley JE: **A critical examination of stoichiometric and path-finding approaches to metabolic pathways**. *Briefings in bioinformatics* 2008, **9**:422-436.

145. Kim PJ, Lee DY, Kim TY, Lee KH, Jeong H, Lee SY, Park S: **Metabolite essentiality elucidates robustness of Escherichia coli metabolism**. *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**:13638-13642.

146. Mahadevan R, Edwards JS, Doyle FJ: **Dynamic flux balance analysis of diauxic growth in Escherichia coli**. *Biophysical journal* 2002, **83**:1331-1340.

147. Price ND, Famili I, Beard DA, Palsson BO: **Extreme pathways and Kirchhoff's second law**. *Biophysical journal* 2002, **83**:2879-2882.

148. Wiback SJ, Mahadevan R, Palsson BO: **Reconstructing metabolic flux vectors from extreme pathways: defining the alpha-spectrum**. *Journal of theoretical biology* 2003, **224**:313-324.

149. Schuster S, Dandekar T, Fell DA: **Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering**. *Trends in Biotechnology* 1999, **17**:53-60.

150. Cornelius SP, Lee JS, Motter AE: **Dispensability of Escherichia coli's latent pathways**. *Proceedings of the National Academy of Sciences of the United States of America* 2011, **108**:3124-3129.

151. Conway T: **The Entner-Doudoroff pathway: history, physiology and molecular biology**. *Fems Microbiology Reviews* 1992, **9**:1-27.

152. Papagianni M: **Recent advances in engineering the central carbon metabolism of industrially important bacteria**. *Microbial cell factories* 2012, **11**:50.

153. Melendez-Hevia E, Waddell TG, Heinrich R, Montero F: **Theoretical approaches to the evolutionary optimization of glycolysis--chemical analysis**. *European journal of biochemistry / FEBS* 1997, **244**:527-543.

154. Nogales J, Palsson BO, Thiele I: **A genome-scale metabolic reconstruction of Pseudomonas putida KT2440: iJN746 as a cell factory**. *Bmc Systems Biology* 2008, **2**:79.

155. Sohn SB, Kim TY, Park JM, Lee SY: **In silico genome-scale metabolic analysis of Pseudomonas putida KT2440 for polyhydroxyalkanoate synthesis, degradation of aromatics and anaerobic survival**. *Biotechnology journal* 2010, **5**:739-750.

156. Puchalka J, Oberhardt MA, Godinho M, Bielecka A, Regenhardt D, Timmis KN, Papin JA, Martins dos Santos VA: **Genome-scale reconstruction and analysis of the Pseudomonas putida KT2440 metabolic network facilitates applications in biotechnology**. *Plos Computational Biology* 2008, **4**:e1000210.

157. Thiele I, Swainston N, Fleming RM, Hoppe A, Sahoo S, Aurich MK, Haraldsdottir H, Mo ML, Rolfsson O, Stobbe MD, et al.: **A community-driven global reconstruction of human metabolism**. *Nature Biotechnology* 2013, **31**:419-425.

158. Hao T, Ma HW, Zhao XM, Goryanin I: **Compartmentalization of the Edinburgh Human Metabolic Network**. *Bmc Bioinformatics* 2010, **11**:393.

159. Agren R, Bordel S, Mardinoglu A, Pornputtapong N, Nookaew I, Nielsen J: **Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT**. *Plos Computational Biology* 2012, **8**:e1002518.

160. Romero P, Wagg J, Green ML, Kaiser D, Krummenacker M, Karp PD: **Computational prediction of human metabolic pathways from the complete human genome**. *Genome Biology* 2005, **6**:R2.

161. Hua Q, Yang C, Oshima T, Mori H, Shimizu K: **Analysis of gene expression in Escherichia coli in response to changes of growth-limiting nutrient in chemostat cultures**. *Applied and Environmental Microbiology* 2004, **70**:2354-2366.

162. Slavov N, Botstein D: **Coupling among growth rate response, metabolic cycle, and cell division cycle in yeast**. *Molecular Biology of the Cell* 2011, **22**:1997-2009.

163. Bennett MR, Pang WL, Ostroff NA, Baumgartner BL, Nayak S, Tsimring LS, Hasty J: **Metabolic gene regulation in a dynamically changing environment**. *Nature* 2008, **454**:1119-1122.

164. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M: **KEGG as a reference resource for gene and protein annotation**. *Nucleic acids research* 2016, **44**:D457-D462.

165. Orth JD, Palsson B: **Gap-filling analysis of the iJO1366 Escherichia coli metabolic network reconstruction for discovery of metabolic functions**. *Bmc Systems Biology* 2012, **6**:30.

166. Satish Kumar V, Dasika MS, Maranas CD: **Optimization based automated curation of metabolic reconstructions**. *Bmc Bioinformatics* 2007, **8**:212.

167. Chen L, Vitkup D: **Predicting genes for orphan metabolic activities using phylogenetic profiles**. *Genome Biology* 2006, **7**:R17.

168. Long MR, Ong WK, Reed JL: **Computational methods in metabolic engineering for strain design**. *Current opinion in biotechnology* 2015, **34**:135-141.

169. Carbonell P, Planson AG, Fichera D, Faulon JL: **A retrosynthetic biology approach to metabolic pathway design for therapeutic production**. *Bmc Systems Biology* 2011, **5**:122.

170. Rodrigo G, Carrera J, Prather KJ, Jaramillo A: **DESHARKY: automatic design of metabolic pathways for optimal cell growth**. *Bioinformatics* 2008, **24**:2554-2556.

171. Yim H, Haselbeck R, Niu W, Pujol-Baxley C, Burgard A, Boldt J, Khandurina J, Trawick JD, Osterhout RE, Stephen R, et al.: **Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol**. *Nature chemical biology* 2011, **7**:445-452.

172. Campodonico MA, Andrews BA, Asenjo JA, Palsson BO, Feist AM: **Generation of an atlas for commodity chemical production in Escherichia coli and a novel**

pathway prediction algorithm, GEM-Path. *Metabolic engineering* 2014, **25**:140-158.

173. Finley SD, Broadbelt LJ, Hatzimanikatis V: **Computational framework for predictive biodegradation**. *Biotechnol Bioeng* 2009, **104**:1086-1097.

174. Hadadi N, Hatzimanikatis V: **Design of computational retrobiosynthesis tools for the design of de novo synthetic pathways**. *Current opinion in chemical biology* 2015, **28**:99-104.

175. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA, et al.: **PubChem Substance and Compound databases**. *Nucleic acids research* 2016, **44**:D1202-1213.

Meriç Ataman

Address : Rue du Maupas 42B, 1004 Lausanne, Switzerland
Tel : +41 78 904 88 28
e-mail: meric.ataman@epfl.ch

SUMMARY

- PhD in Systems Biotechnology, MSc in Chemical and Biological Engineering
- Background in Computational Biology, Metabolic Modeling and Engineering
- Recognized Abilities in Presentation Skills and Communication

EDUCATION

- **2011- : PhD in Chemistry and Chemical Engineering**, **Ecole Polytechnique Fédérale de Lausanne**, LAUSANNE, SWITZERLAND, Laboratory of Computational Systems Biotechnology
- **2009-2011: MSc in Computational Biology**, **Koc University,** Department of Chemical and Biological Engineering, ISTANBUL, TURKEY, Systems Lab
- **2004-2009: Chemical Engineering**, **Bogazici University**, Faculty of Engineering, ISTANBUL, TURKEY

PROFESSIONAL EXPERIENCE

November 2014 – April 2015: Research Intern, Nestlé Institute of Health Sciences (NIHS)
• Reconstruction of a Genome Scale Metabolic Network

- Analyses on growth and by-product formation on various media
- Thermodynamics curation of Genome Scale Metabolic Network

Research Assistant and PhD Candidate, **EPFL**

• Using constrained based methods for understanding cellular metabolism with a systems biology perspective.

• Developing computational methods to discover the metabolic capacity of model microorganisms.

• Providing consultancies for various industrial projects for the identification of novel biosynthesis pathways for the production of value-added chemicals and biofuel

• Teaching assistant for “Introduction to Chemical Engineering”, “Principles and Applications of Systems Biology” and “Biochemical Engineering“ courses and supervising bachelor and master level students

COMPUTER SKILLS

Operating Systems: Mac OS, Windows, LINUX

Programming: MATLAB, C

Process simulators: ChemCAD

Productivity Software: Microsoft Office (Word, Excel, Powerpoint)

PRESENTATIONS AND PUBLICATIONS

- Ataman M, Hatzimanikatis V: **Heading in the right direction: thermodynamics-based network analysis and pathway engineering**. *Current opinion in biotechnology* 2015, **36**:176-182.
- Hadadi N, Ataman M, Hatzimanikatis V, Panayiotou C: **Molecular thermodynamics of metabolism: quantum thermochemical calculations for key metabolites**. *Physical chemistry chemical physics : PCCP* 2015, **17**:10438-10453.
- Merten Morales, Meriç Ataman, Sara Badr, Sven Linster, Ioannis Kourlimpinis, Stavros Papadokonstantakis, Vassily Hatzmanikatis and Konrad Hungerbühler**: Sustainability Assessment of Succinic Acid Production Technologies from Biomass using Metabolic Engineering,** *Energy Environ. Sci.*, 2016
- Meriç Ataman, Daniel Hernandez, Georgios Fengos, Keng Cher Soh, Vassily Hatzimanikatis: **"*redGEM*: A unbiased approach for systematic reduction of genome-scale models"** 251st ACS National Meeting & Exposition March 13-17, 2016, San Diego, California, USA. Oral Presentation

LANGUAGES

• Turkish (mother tongue) • English (full working proficiency) •French (beginner, A2)

HOBBIES

• Music, playing guitar, history, politics, writing short stories, football, beach volley, whiskey