Computational Studies on Cellular Metabolism: From Biochemical Pathways to Complex Metabolic Networks

THÈSE N^{O} 6667 (2015)

PRÉSENTÉE LE 11 SEPTEMBRE 2015

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

Noushin HADADI

acceptée sur proposition du jury:

Prof. U. Röthlisberger, présidente du jury
Prof. V. Hatzimanikatis, directeur de thèse
Prof. L. Blank, rapporteur
Prof. H. Riezman, rapporteur
Prof. G. Vandergoot, rapporteuse

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2015

ACKNOWLEDGMENT

In the past five years, I had invaluable support and guidance from many people that I will forever be thankful to them.
First and foremost, I would like to express my sincerest gratitude to my enthusiastic and truly dedicated supervisor and mentor, Vassily Hatzimanikatis, a person that you will never forget once you meet him. He is one of the smartest, most generous and lively people I know. I admire his passion for science and I wished several times that the curiosity and enthusiasm he has for science and research was contagious! Thank you Vassily for always being supportive, patient, caring and encouraging with me throughout my PhD education. I've learned a lot from you and I am indebted to you not only for outstanding academic support and so many life advises, but particularly for your trust and belief on me and my work. I will miss the wonderful tea and lunchtime we had and the smile that your humors and jokes bring to my face, especially in the situations that I would get unreasonably serious!
I take this opportunity to thank the members of my PhD defense committee for having accepted to assess my thesis, for their encouraging words and valuable feedbacks: Prof. Ursula Roethlisberger, Prof. Howard Riezman, Prof. Lars Blank and Prof. Gisou. van der Goot. I would like to especially thank Gisou for being encouraging from the very beginning of my Ph.D. You are the excellent example of a successful woman and scientist for me.
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.

The present and past members of the LCSB group have contributed immeasurably to my personal and professional time at EPFL. First, I would like to say a very big thank you to Dr. Ljubisa Miskovic for being my primary resource of help and for his scientific and moral support during the past five years. Thank you Misko for being so generous on helping no matter the task or condition, for listening to my complaints and always giving me the best advice and for so many insightful discussions we had. Your continuous support and advise during this period played a major role on the fulfillment of this thesis.
Special mention goes to Meric Ataman, my collaborator on many industrial projects and who has contributed to the work presented in this thesis. It has been a pleasure to work with you Meric!

My deep appreciation goes out to my office mates for making the time working on my Ph.D. an unforgettable experience. Andrijana, you are a real friend! Katerina and Stepan, I have very fond memories of the time we shared together in the office and I still get a laugh when I remember so many wonderful memories we have from that time! Katerina, or as I call you: my dearest Katerina, in you I have a life-long friend and I can not express my gratitude and feelings for you. And my last office mate, Pierre, thank you for being so understanding in the last months while I was writing my thesis and was a bit! irritable and of course for all your helps in French!
I thank other past and present group members for their support and for providing a pleasant and productive working atmosphere: Keng for his constant help on different projects, Marianne, Anirikh, Julien, Stefano, Alex, Yves, Georgios and Georgios! and Vikash for answering all my questions with smiley faces. Joana, my hair-care consultant for all the nice discussions we had! Tiziano, Daniel, Milenko, Tuure, and Anush whom I think of as a little sister, for bringing the nice cheerful atmosphere to the lab!
I'm also glad to have worked with many undergraduate students: Pamela, Jake, Sadegh, Alexendre and Christian. Special thanks go to Jasmin and Adrian, for the hard work and their great contribution in this thesis.
Being far from my home country for 7 years of Master and PhD studies, I am indebted to several Iranian friends who helped me wherever I needed and enriched my time at EPFL. Mona, you are a wonderful and generous friend and I'll never forget those many days that we laughed and cried together and all the life-matter discussions we had! I was lucky to meet Atefeh, my first true friend in Switzerland, Nina, Laleh and Faezeh and have them as close friends in the past years; your friendship means a lot to me!

I profoundly thank my parents for raising me with infinite love, for their unwavering belief in me and for always encouraging me to follow my dreams. I want you to know how much I appreciate the life lessons I got from you about self-respect, resilience, strength and how to be independent and am forever indebted to you. I want to thank my sister, Mehrnoush, whom I miss everyday, for all her encouragements and life advises through every day phone calls; thousands of kilometers couldn't take us apart! My little sister Roshanak is the most strong and amazing person I ever met in my life; I'm so proud of you Roshi and I'm grateful for all your unconditional care and help, whenever I needed it. And my little brother Sasan, who is always there to help and to listen. I wish you understand one day how much your tough sister loves you.
Lastly, my deepest thanks and appreciations go to my biggest supporter, my husband, Pedram. I've been able to complete this long dissertation journey because I had you always by my side, from the beginning to the end, during my good and bad times. Your love and the faith that you have in me make me feel like there is nothing in this world that I cannot accomplish. These past several months have been a challenging period for me, and you, the most patient person I know, went through all the mood changes and frustrations with me. There are no words that can express my gratitude for all you've done and been for me and for always trying to see the best of me.

ABSTRACT

Biotechnology promises the biologically and ecologically sustainable production of commodity chemicals, biofuels, pharmaceuticals and other high-value products using industrial platform microorganisms. Metabolic engineering plays a key role in this process, providing the tools for targeted modifications of microbial metabolism to create efficient microbial cell factories that convert low value substrates to value-added chemicals. Engineering microbes for the bioproduction of chemicals has been practiced through three different approaches: (i) optimization of native pathways of a host organism; (ii) incorporation of heterologous pathways in an amenable organism; and finally (iii) design and introduction of synthetic pathways in an organism. So far, the progress that has been made in the biosynthesis of chemicals was mostly achieved using the first two approaches. Nevertheless, many novel biosynthetic pathways for the production of native and non-native compounds that have potential to provide near-theoretical yields and high specific production rates of chemicals remain yet to be discovered. Therefore, the third approach is crucial for the advancement of bio-based production of value-added chemicals.

We need to fully comprehend and analyze the existing knowledge of metabolism in order to generate new hypotheses and design *de novo* pathways. Current knowledge about the metabolism is incomplete even for well-studied microorganism such as *E. coli* and *S. cerevisiae*, and researchers keep discovering and characterizing new biochemistry, genes, enzymes and reaction mechanisms. The complexity of metabolic networks cannot be understood and analyzed intuitively; thus, computational methods are indispensable to explore and expand the present knowledge of metabolism in order to unravel the full enzymatic potential of microorganisms.

In this thesis, through development and application of efficient computational methods, we took the research path to expand our understanding of cell metabolism with the aim to discover novel knowledge about metabolic networks. We analyze different aspects of metabolism through five distinct studies.

In the first study, we begin with a holistic view of the enzymatic reactions across all the species, and we propose a computational approach for identifying all the theoretically possible enzymatic reactions based on the known biochemistry. We organize our results in a web-based database called “Atlas of biochemistry”.

In the second study, we focus on one of the most structurally diverse and ubiquitous constituents of metabolism, the lipid metabolism. Here we propose a computational framework for integrating lipid species with unknown metabolic/catabolic pathways

into metabolic networks. In our next study, we investigate the full metabolic capacity of *E. coli*. We explore computationally all enzymatic potentials of this organism, and we introduce the "Super *E. coli*", a new and advanced chassis for metabolic engineering studies.

Our next contribution concentrates on the development of a new method for the atom-level description of metabolic networks. We demonstrate the significance of our approach through the reconstruction of atom-level map of the *E. coli* central metabolism and we show how a comprehensive atom-level metabolic analysis can guide the experimental design to obtain more precise biological information.

In the last study, we turn our focus on studying the thermodynamics of metabolism and we present our original approach for estimating the thermodynamic properties of an important class of metabolites. So far, the available thermodynamic properties either from experiments or the computational methods are estimated with respect to the standard conditions, which are different from typical biological conditions. Our proposed workflow paves the way for reliable computing of thermochemical properties of biomolecules at biological conditions of temperature and pressure.

Finally, in the conclusion chapter, we discuss the outlook of this work and the potential further applications of the computational methods that were developed in this thesis.

Keywords: Metabolism, metabolic networks, systems biology, metabolic engineering, computational biochemistry, computational biology, *de novo* pathway design, thermodynamics

RÉSUMÉ

Les biotechnologies portent la promesse d'une production biologiquement durable et responsable de composés chimiques de base, biocarburants, médicaments ou autres produits à haute valeur ajoutée, grâce à l'utilisation de micro-organismes comme plateforme industrielle. L'ingénierie métabolique joue un rôle clef dans ce schéma, puisqu'elle fournit les outils nécessaires à une modification ciblée du métabolisme des micro-organismes afin de créer des usines microbiennes capables de convertir un substrat primaire en un produit chimique à haute valeur ajoutée. Actuellement, on observe l'utilisation de trois approches pour modifier un micro-organisme et lui faire produire une molécule d'intérêt : (i) L'optimisation des voies métaboliques préexistantes chez l'hôte, (ii) l'incorporation de voies métaboliques appartenant à un autre organisme, (iii) la construction et l'introduction de voies métaboliques synthétiques dans ledit organisme. Jusqu'à maintenant, les deux premières méthodes ont été les plus explorées et utilisées. Toutefois, il existe encore de nombreux composés pour lesquels il n'y a pas de voie métabolique connue présentant hauts rendement et productivité. C'est pourquoi la troisième approche, la construction de voies synthétiques, reste primordiale afin d'assurer le développement de la bio-production de produits chimiques à haute valeur ajoutée.

Afin de créer et construire de nouvelles voies métaboliques, il est nécessaire d'avoir une compréhension exhaustive du métabolisme. Pour le moment, même des organismes pourtant très étudiés comme *Saccharomyces Cerevisiae* ou *Escherichia Coli* présentent des lacunes dans leurs modèles métaboliques, et les chercheurs continuent encore de découvrir et caractériser de nouveaux gènes, enzymes, interactions biochimiques et mécanismes réactionnels. La complexité des réseaux métaboliques n'est pas intuitive, et c'est pourquoi des méthodes computationnelles sont indispensables pour l'explorer et améliorer la connaissance actuelle du métabolisme en général, afin d'être en mesure de maximiser l'impact du potentiel enzymatique des micro-organismes.

Dans cette thèse, à l'aide du développement et de l'application de méthodes computationnelles de pointe, j'ai oriente ma recherche vers la compréhension du métabolisme cellulaire, dans le but d'en découvrir plus sur les réseaux métaboliques. J'analyse différents aspects du métabolisme via cinq études distinctes.

La première adopte un point de vue holistique, et s'intéresse à l'ensemble des réactions enzymatiques connues. J'y propose une approche computationnelle permettant d'identifier toutes les réactions enzymatiques théoriquement possibles, à partir des

interactions biochimiques déjà connues. Les résultats sont organisés dans une base de donnée appelée *Atlas de la biochimie*.
La seconde étude s'intéresse à un des systèmes les plus structurellement versatiles, et néanmoins indispensables, du métabolisme : le métabolisme des lipides. J'y propose une méthodologie computationnelle permettant d'intégrer dans un réseau métabolique des lipides n'ayant pas de voie métabolique ou catabolique connue.
Le chapitre suivant porte sur l'exploration de la totalité de la capacité métabolique de *Escherichia Coli*. J'y recense tout le potentiel enzymatique de cet organisme, et y introduis « *Super E. Coli* », un nouveau châssis servant de cadre d'étude pour l'ingénierie métabolique de cet organisme.
La pénultième contribution s'attache au développement d'une méthode de description atomique des réseaux métaboliques. J'y démontre la pertinence de mon approche en reconstruisant une carte métabolique à l'échelle atomique du métabolisme central du carbone chez *Escherichia Coli*, et explique comment une analyse métabolique exhaustive à cette échelle peut guider la conception de nouveaux protocoles expérimentaux visant à obtenir des données biologiques plus détaillées
Enfin, La dernière étude s'intéresse à l'étude thermodynamique du métabolisme, et j'y présente une approche originale estimant les propriétés thermodynamiques d'un grand nombre de métabolites. Jusqu'à présent, les données thermodynamiques provenant d'expériences ou bien de méthodes computationnelles sont estimées dans l'état thermodynamique standard, qui est typiquement différent des conditions cellulaires. La méthodologie présentée établit les bases d'un calcul précis des propriétés thermochimiques des métabolites, dans des conditions de pression et température physiologiques.
En conclusion, je récapitule les implications de ces différents projets, et élabore sur les applications potentielles des méthodes computationnelles développées dans cette thèse.

Mots-clefs : Métabolisme, réseaux métaboliques, biologie des systèmes, ingénierie métabolique, biochimie computationnelle, biologie computationnelle, conception de voies métaboliques *de novo*, Thermodynamique.

CONTENTS

LIST of FIGURES

LIST of TABLES

TABLES IN THE APPENDIX

INTRODUCTION

Motivation

For the first time ever, half of the people living on earth are urban residents and by 2050, it is expected that 70% of the global population will live in the cities [1]. Rural to urban migration brings forward new challenges relevant to the urban population growth in the future. Urban communities are facing two big challenges: (i) limited energy and food resources and (ii) unsustainable consumption of these resources, which result in increasing waste streams production and CO_2 emissions. These are two of the most critical issues for the next generations and are tightly linked to the fact that the world's population is extremely fossil fuel dependent [2].

The creation of sustainable "white" cities would be the ultimate solution and is pictured as: cities that exploit the benefits of *white biotechnology*.

We can imagine future urban habitations where the city powers itself with renewable sources of energy, recycles material at various levels (food, clothing, objects of daily use, etc.), converts non-recyclable waste to energy, and thereby reduces the ecological footprint.

White biotechnology promises the biologically and ecologically sustainable production of valuable chemicals, fuels and other high value products using "modified" microorganisms known as Synthetic Microbial Cell Factories (SMCFs), which serve as a substitution to the unsustainable petrochemical-based processes [3,4]. These robust biological machines are designed as revolutionary platforms to produce valuable chemicals and biofuels (to address the limited sources of fossil fuels) using sustainable resources and waste biomass (to address environmental concerns). Therefore, SMCFs are the ideal means for building and supporting white cities through closing the loop cycle of waste being turned into valuable products and sustainable energy sources.

However, most of natural organisms cannot directly serve as MCF to produce desired compounds or to decompose man-made material used in modern lifestyle. Therefore,

prior to their application as MCFs, the platform microorganisms need to be adapted and genetically modified through retrofitting and optimization processes introduced in *synthetic biology* and *metabolic engineering* disciplines. The very important question is how to manipulate and harness the metabolic capabilities of organisms.

To do such modifications through metabolic engineering, one should have a complete understanding of *metabolism* across different organisms and how metabolic processes determine and regulate the physiological and biochemical properties of a cell. Metabolism comprises all the biochemical transformations that occur in living organisms, and is organized through interconnected metabolic and regulatory networks.

Metabolism is extremely complex and requests a *systems approach* towards understanding the involved metabolic processes for delivering solutions for metabolic engineering applications. With the emergence and current progress made in high throughput technologies, great quantities of biochemical and biological data are generated in genomics, metabolomics and proteomics studies that help elucidate the constituents of metabolic and regulatory networks and their interactions.

Mathematical approaches are proven to be imperative for organizing and integrating the generated "*omics*" data into metabolic models and to enhance our understanding of metabolism and uncovering its complexity. Mathematical descriptions of metabolism (metabolic models) allow a systematic compilation of the generated *omics* data of an organism and their analysis allows acquiring significant knowledge about the entire metabolic network. Understanding the metabolism is not only crucial for the success of metabolic engineering through design and delivering microbial platforms for white biotechnology, but also has important implications for discovering novel therapeutic approaches by unraveling the complicated mechanisms leading to diseases where metabolism plays a critical role.

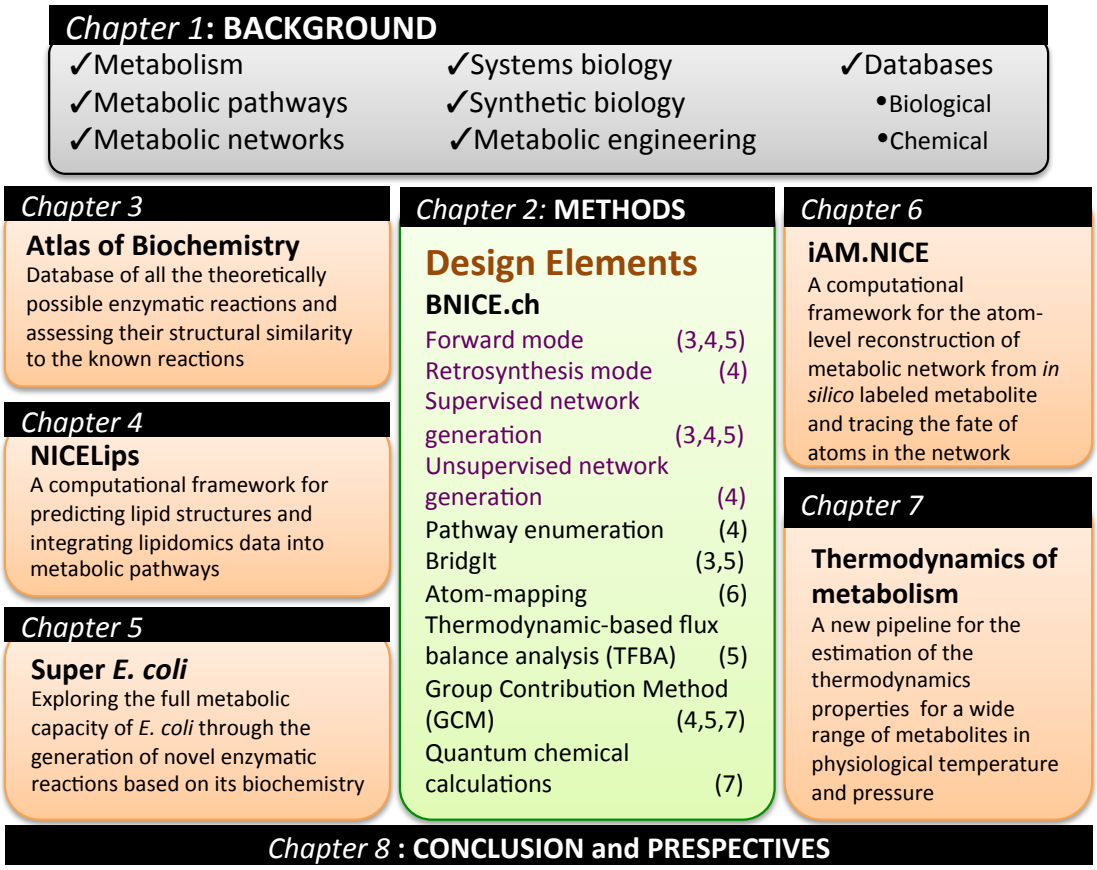
The aim of this thesis is the development of a suite of computational methods to investigate metabolic networks and to gain novel biological perceptions of metabolism. We apply our methods in a diverse range of problems in systems biology. Doing so, we demonstrate how computational approaches coupled with the engineering fundamentals of process design can explore and expand the potential of metabolism. As

such, the outcome of this thesis highlights the importance of computational approaches in accelerating metabolic engineering studies and in providing guidance for designing therapeutic interventions.

Thesis Structure

This thesis is organized into nine chapters. Following the current introduction, in the following chapters, we describe and discuss our approaches for addressing these challenges. We provide a graphical description of the thesis structure that gives an overview of the different chapters of this dissertation and it lists different tools and design elements we developed during this PhD work, together with the chapters that they have been applied.

In **Chapter 1,** we introduce the definitions and the necessary background for the



essential notions in this thesis. We further briefly discuss the progress made so far and the challenges remained to be addressed in the fields of systems biology and metabolic engineering. **Chapter 2** describes and explains the technical background and the main methodologies underlying the research in this thesis. We comprehensively review and explain various existing computational methods for the analysis of metabolic networks and compare them with our proposed methods.

In the next five chapters, we address five research problems using extensions of our developed methods. These works were published or submitted as research articles and focus on a specific aspect of metabolism. Each chapter starts with a research question that we wanted to answer in that study, and a short summary followed by an introduction of the specific aspects discussed in that work, the findings and results of the work and the concluding remarks of the study.

In **Chapter 3** we present the application of our developed methods for exploring the KEGG reaction database. **Chapter 4** represents the computational studies of lipid metabolism. In **Chapter 5** we incorporate several methods for the development of a comprehensive metabolic network for *E. coli*. In **Chapter 6** we extend one of our developed methods for the atom-mapped reconstruction of metabolic networks with a demonstrative example of *E. coli* core metabolic network. In **Chapter 7** we introduce our pipeline to study the thermodynamics of metabolism.
Finally, **Chapter 8** concludes this thesis and gives some perspectives for future researches based on the findings of this work.

Chapter 1

BACKGROUND

In this chapter we review essential biological and chemical notions and definitions important for this thesis including metabolism and its relevant subjects such as metabolic pathways, metabolic network and metabolic models. We further focus our attention on the lipid metabolism and its significant roles in most of the biological systems and its association with several diseases. An introduction to atom transition in metabolism and the bioenergetics of metabolism are followed.

We also briefly present the emerging research fields of systems biology, synthetic biology and metabolic engineering that rely on the understanding of metabolism and for which the outcome of this thesis would have several applications. Finally we introduce and explain the importance of available biological and chemical databases and their organizations.

1.1 Metabolism, from metabolic pathways to metabolic networks

Metabolism is the ensemble of biological reactions that consecutively interconvert metabolites in the living cells to sustain life. These biochemical reactions are catalyzed by specialized proteins, called enzymes, which convert nutrient molecules to biomass building blocks and release energy to power different biochemical processes in the cell needed for growth [5]. Metabolism is one of the most complex cellular processes and understanding it is crucial to many disciplines such as systems biology, biotechnology, medicine and pharmaceutical research for designing new drugs [6].

For facilitating the study of metabolism, researchers usually divide it into two

categories: catabolism and anabolism. They further categorize it into smaller subunits called *metabolic pathways* [7].

Metabolic pathways are a series of enzyme-catalyzed reaction steps which convert one compound into another through catabolic or anabolic ways. In the case of breaking down and oxidizing the large molecules (nutrient and etc.) into small substrates, this process is called catabolism. *Catabolic pathways* deliver the energy and substrates needed for anabolism. *Anabolic pathways* make use of the released energy and small molecules from catabolic reactions to produce molecules that construct the cell components such as proteins, lipids and amino acids [8].

Traditionally, biologist studied metabolism by focusing on a specific metabolic pathway and investigated the different involved reaction steps and enzymes, using different approaches such as tracer experiments [9]. These analyses provided significant insights into metabolism and brought a clear understanding of the structure and the function of canonical metabolic pathways, such as glycolysis (the first discovered metabolic pathway), pentose phosphate pathway, citric acid cycle, etc.

However, it is now believed that the metabolic pathways do not function as individual isolated constituents and that their simultaneous function and interactions result in an extremely connected complex network [10]. Hence, the historical reductionist and simplified approaches cannot address numerous questions arising from the study of the whole cell where the interactions of different parts create the emergent properties and behaviors. The plethora of all compounds, reactions, enzymes and metabolic pathways data along with their regulatory interactions which would determine the physiological and biochemical properties of a cell are integrated into *metabolic networks*. Metabolic networks are the backbone of *metabolic models* and are fundamental for metabolism modeling and simulations [11].

Following the advances in genome sequencing techniques, the complete genome sequence is becoming available for numerous organisms, and thus the genomic data can be integrated in the metabolic models. Such assimilation results in the generation of more comprehensive metabolic models that not only explain the physiological properties of cell, but also can predict their behaviors. We elaborate more on the different modeling approaches for the metabolic network analysis and their significance

in Chapter 2.

1.2 Lipid metabolism

Lipids are hydrophobic organic molecules that include some common compounds such as fats, oils, waxes, phospholipids and steroids (e.g., cholesterol). In addition, they include a large range of compounds with special structures and functions. Lipids have several key biological functions in living cells such as participating in signaling pathways, acting as energy storage sources and being the structural components of cell membrane. Moreover, they play an important role in various diseases such as obesity, diabetes and cancer [12]. Lipid-associated disorders are metabolic disorders and their characteristic is the accumulation of intolerable amounts of lipids in some of the body's cells as a reason of defects in the biosynthesis or biodegradations of simple lipids [13,14].

The term lipidome refers to the full lipid complement of cells, tissues and organisms. Lipidomics – the large-scale study of pathways and networks of cellular lipids in biological systems – aims to elucidate and characterize the lipidome [15,16]. Mass spectrometry (MS) is the most commonly used analytical method in lipidomics research [17-19]. With the rapid growth in analytical technologies, in particular MS, vast amount of data is being generated for lipid structures [20]. Therefore, there is a need for developing comprehensive computational tools for data mining and system level identification of lipid species and organizing them into databases [21].

Although there are a number of ongoing discoveries in lipidomics, assisted by functional genomics and biophysical studies, still a large number of structures and functions need to be discovered or clarified in lipid metabolism [22,23]. Due to the diversity of possible chemical structures in lipids, bioinformatics is indispensable for accelerating the discovery of novel structures and their corresponding biosynthetic and catabolic pathways. In Chapter 5, we present our computational approach for the analysis of glycerophospholipids metabolism, a diverse and ubiquitous class of lipids and the results of a case study to show the applicability of our methods for lipid research.

1.3 Atom transition in metabolism

Biological pathways have been extensively analyzed with respect to overall conversions of substrates to products. However, information on atom transitions in metabolic pathways is not widely available in the literature [24]. Reaction atom mappings track the positional changes of all the atoms between the substrates and the products as they undergo the chemical transformation and elucidate the mass flow in a biochemical reaction network [25]. The understanding of metabolic reactions at the atomic level is of great importance as it can deconvolute the overlapping catabolic/anabolic pathways resulting in the observed metabolite [26]. This has widespread impact on applications of systems biology such as isotope labelling experiments, pathway inference in genome reconstructions, flux quantification for metabolic engineering and strain development, and in the studies of metabolic diseases [27,28].

There are two main approaches to study the atom transition through a reaction: experimental approach versus *in silico* approach.

Determining the atom map of a reaction *experimentally* means to replace one or several atoms with radio-isotopes or, more recently, stable-isotopes. Due to the difference in the mass of isotopes, the labeling pattern of the product can be determined by mass spectrometry. In contrast, *in silico* atom mappings require no isotopes. Instead, they require chemical/biochemical knowledge about the reaction mechanism. *In silico*, or computational, methods have been developed to automatically generate atom maps for large sets of reactions. Still, isotope-labeled metabolites are the standard approach to elucidate reaction mechanisms. The computational determination of reaction mechanisms, however, would be more efficient and would help to reduce the amount of experimental work.

^{13}C Metabolic Flux Analysis (^{13}C-MFA) [29-31] is the standard approach to elucidate the reaction mechanisms as well as the particular distribution of fluxes in an organism. However, isotopic labelling experiments are costly and time-consuming. The computational determination of reaction mechanisms is more efficient and can guide experimental design. ^{13}C-MFA techniques require a description of atom transitions from substrates to products for each reaction in order to make predictions [32]. The ability to track atoms through reaction pathways is inevitable in ^{13}C-MFA studies[29,33] in order

to: (i) find the appropriate position of ^{13}C in the tracer metabolite; and (ii) choose the best metabolites or biomass components to be measured in order to maximize the information content of flux observations in the metabolic network [34].

In a comprehensive *in silico* atom-level reconstruction of a metabolic network, the atom mappings need to be: (i) correctly created at the level of individual reactions, and (ii) connected in a network and conserved through all reactions steps from the input compound to the final products, enabling one to trace back the exact metabolic path of every single atom.

More discussions about the computational methods for atom mapping and our contribution to the field are provided in Chapter 2 and we further demonstrate the results of our automated atom mapping method in Chapter 6.

1.4 Bioenergetics of metabolism

Bioenergetics is the crucial aspect of the metabolic processes as it concerns the energy production and transformations in biological systems [35].

Bioenergetics is involved with any bonds that are broken or made in the molecules, which are part of the cellular processes. As already mentioned, metabolism is divided into two categories, and in catabolism, the breaking of nutrients into smaller components leads to the production of energy in different forms such as adenosine triphosphate (ATP) molecules. The released energy powers anabolism to construct complex molecules that serve as biomass building blocks needed for growth. Life depends on this energy flow and understanding the detailed mechanisms involved in the energy transformation in living organisms is a big challenge [36].

The laws of thermodynamics hold true in the metabolism of a cell. Biological thermodynamics concerns the chemical thermodynamics in biological and biochemical systems. In thermodynamics, the energy required or released in the course of a chemical reaction is measured quantitatively by the change in the Gibbs free energy (Δ_rG). Δ_rG determines the degree of thermodynamic favorability of a reaction, *i.e.,* a negative Δ_rG indicates that the reaction can occur spontaneously, and if Δ_rG is positive, the reaction is nonspontaneous in the designated direction [37].

Thermodynamics has been applied in the context of biological systems to improve our understanding of the metabolism. It has been used to quantify the feasible ranges for the Gibbs free energy change of a reaction [38-40] and to systematically assess the degree of reversibility of metabolic reactions [41]. Such analysis requires a sufficiently accurate knowledge of the thermochemical quantities of involved substrates and metabolites [42]. The number of molecules involved in the biochemical reactions is huge and the experimental available Gibbs free energies of formation cover only a small fraction of them [43,44]. Hence, due to the scarce amount of available experimental data for the biological systems, experimental-based thermodynamics analysis of metabolism would be limited to small subsections of metabolism. Therefore, methods for the efficient use of this minimal information and the development of reliable predictive schemes of the thermochemical quantities are compulsory to fill in the gaps in the experimental data [45-47].

Group-Contribution method (GCM), a particularly useful tool for metabolic pathway analysis, is developed in Hatzimanikatis lab in 2008 to address the issues of the limited experimental information for the thermodynamics of biological systems.

In this method, a molecule is decomposed into their constituent groups of atoms and the Gibbs energy of formation of the compound can be estimated as the summation of the properties of groups [48].

Using GCM, the standard Gibbs free energy of formation (and thus, the corresponding free energy change of their reactions) can be estimated with sufficient accuracy for a large percentage of dilute aqueous solutions [48]. GCMs are indispensable tools for the high throughput feasibility analysis of metabolites and pathways. However, since they are empirical methods, they do have their limitations. The coverage and the quality of functional group contributions heavily depend on the existence of relevant experimental data and on the quality of the data used for their determination. Thus, the development of alternative non-empirical approaches, which could substitute experiments, are crucial not only for the practical calculations but also for the insight they provide. Such developments can also be used to enrich and improve GCMs or to help with consistency tests. With today's advances in *ab initio* and density functional theory (dft) calculations and the availability of computational resources with ever

increasing speed, the development of the above mentioned alternative non-empirical approaches through *quantum chemical calculations* appear feasible.

Another consideration about the experimental approaches and consequently GCM is that data - either obtained experimentally or estimated through GCM – are for standard conditions (298.15 K, 101 kPa), which is not the case in most biological systems. Adjusting the obtained data from GCM to the cellular conditions is therefore an essential step for the improvement of the results obtained through this method. Knowledge about the thermodynamics properties at non-standard conditions is of a great interest for many applications in biology, biotechnology and medicine such as the following.

1. The temperature of human body could change from 37^0C in a healthy condition to 40^0C in a sick condition. For understanding and exploring the biochemistry and molecular biology of human metabolic networks in both healthy and especially sick conditions, one needs to have the thermodynamic properties at biological conditions rather than standard.
2. There is evidence that the metabolic activity of gut microbiota has a direct effect on human health. The rate of microbial activity strongly depends on the temperature, and it has been shown that different microbial communities are adapted to operate under different optimal temperature regimes. The availability of thermodynamic properties under relevant temperature for bacterial communities would be essential to understand the underlying mechanisms of microbiota and host interactions.
3. Ancient deep-sea hydrothermal systems provide an environment conducive to the abiotic synthesis of biomolecules that are essential for the emergence of life. The enormous biotechnological potential in harnessing the metabolic capabilities of these extremophiles can only be used with a better understanding of the metabolic requirements and energetics of microbial life at the edges of survivability [49]. In this respect, reliable calculations of thermochemical properties of metabolites and biomolecules at high temperatures and/or high pressures are of primordial importance.

To overcome the limitation of experimental-based thermodynamics and the GCM, one needs to improve the GCM method to obtain thermodynamic predictions that are adjustable to any temperature and pressure based on the nature of the study. In Chapter 2, we discuss the available methodologies and our recently introduced

framework for estimating the thermodynamic properties for a broader range of metabolites in biological systems, respecting the biological conditions of temperature and pressure. Our introduced pipeline takes four main steps to do such estimations, and the results of the first step of the project are provided in Chapter 7. The three other steps are ongoing projects and their results are not included in this thesis.

1.5 Systems biology

The inter-disciplinary research field of *systems biology* aims at the systematic study of metabolic networks through integrating high-throughput '*omics*' experimental data into predictive mathematical models to provide *holistic* views of metabolism [50].

Hence, the pillars of systems biology are the experimental techniques and computational methods to assess the perspective of the entire metabolic network [51].

Opposing to the reductionist approach that study the *snapshots* of metabolism through single biological pathways analysis, systems biology, owing to its network-based intrinsic, takes a systematic holistic approach to study instantaneously all the constituents of complex biological systems along with their interactions.

The considerable advances already made in field of systems biology are due to the fast development and growth of high-throughput measuring technologies such as metabolomics, transcriptomics and proteomics along with the emergence and advances of bioinformatics and computational biology [52].

1.6 Synthetic biology and metabolic engineering

The origins of two emerging fields of *synthetic biology* and *metabolic engineering*, their overlaps and their distinctional disciplines and approaches have been clarified in a recent review by G. Stephanopoulos [53].

While metabolic engineering seeks to customize the cell and pathway performance for the production of desired compounds, synthetic biology pursues this aim through designing and constructing the synthetic genetic circuits that regulate the performance of the cells [53,54]. The new generations of microbial strains, called Synthetic Microbial Cell Factories, enable the sustainable production of a wide range of chemicals and fuels. This demonstrates the achievements being made in the last 20 years by the emergence

and collaboration of multidisciplinary fields of metabolic engineering, systems biology, systems biotechnology and synthetic biology [55,56].

In order to create efficient microbial factories and broaden the range of biosynthetic pathways for the production of both natural and non-natural compounds, it is necessary to go beyond natural pathways by exploring the chemistry and synthetic capability of biological systems [55,56].

Designing an SMCF has several steps, starting with the choice of the target desired compounds and a particular chassis organism. It follows with the further analysis of the biochemistry of the chosen organism to see whether the desired compound is native to the metabolism or not. In case of a non-native compound, metabolic engineering strategies are required to design pathways for the production of the desired compound [57]. After considering several criteria, the best candidate pathways are then engineered in the metabolic network of microorganisms that serve as synthetic platforms in synthetic biology.

The complexity of biological chemistry and metabolism requires computational approaches that explore the full possibility of synthetic pathways towards target compounds. The *de novo* design of pathways is the key to exploit the incredible natural diversity of enzymatic transformations. Detailed discussions about the available methods for pathway design and our contributions to the field are provided in Chapter 2.

1.7 Data organization in biological and chemical databases

In order to take the full benefit from the wealth of extensive and rapidly growing amount of information produced through high-throughput technologies and computational methods, it is crucial to systematically organize and catalogue the generated data into shared central resources to enable storing, searching and retrieving from the data. Last decade has been the successful era of such organized and field specific public data sources. Within the field of systems biology, several specific databases have been developed with different scopes, coverages and prioritizations, and one could broadly classify them to the repository of protein sequences [58,59] or metabolic pathway databases [60,61]. A major challenge is to provide globally

established nomenclatures and representative identities for chemical and biological structures. This would allow connecting different databases in order to compare and browse the data. Lots of efforts have been made to establish such connections between online databases, which make it possible to gather different levels of information based on different needs. In the following chapters we describe the scope and characteristics of the major sources of information available for genes, enzymes, reactions and pathways crucial for metabolic network analysis and reconstruction. These sources have been extensively used in the different research projects throughout this thesis.

Chapter 2

METHODOLOGY & THEORY

In this chapter we review the theory and the technical aspects underlying the research of this thesis. We emphasize on the "*technical design elements*" that were needed to carry out this research and on the *methods* that have been developed and applied to address several biochemical and biological questions in the context of this thesis. We begin with the computational methods for the metabolic network analysis and our contributions for the design of *de novo* pathways. We further introduce the methods we developed for the automated atom mapping of the biological reactions. In the last section, we describe a new pipeline for the improvement of the estimation of the thermodynamic properties. More details about methodologies are covered in the following chapters. After introducing each design element, we point to the chapter in which it is more elaborated later in the thesis.

2.1 Computational methods for metabolic network analysis

Metabolic systems are complex interconnected networks that include genes, proteins, enzymes, metabolites and reactions. By the ever-increasing availability of "*omics*" data, the size and the complexity of metabolic networks are growing [10]. Such complex and large networks cannot be understood and analyzed intuitively; thus, computational methods and strategies for the reconstruction and analysis of metabolic networks are indispensable [10,62]. *Computational metabolic network reconstruction* is a well-established discipline in which we systematically collect the available data to form a knowledge-based mathematical model of the metabolic network for managing the complexity of the biological systems. Metabolic network reconstructions are mostly

organism specific and depend on the quality of the sequenced genome, thus differing in size and quality. If the metabolic network reconstruction is based on the full genome annotation of an organism, we use the term *genome-scale metabolic networks* [63]. These reconstructions collect all the relevant metabolic knowledge of the organisms and capture their metabolic capabilities and provide an in-depth insight into their metabolism.

2.1.1 Genome-scale metabolic models

Genome-scale metabolic models (GEMs) are one of the most important computational resources for the investigation of metabolism and have been constructed and experimentally verified for several organisms. The reconstruction process starts with collecting the relevant knowledge about the given organism from genome analysis, the databases of metabolic information, and its known biochemistry. This process is followed by assembling the collected information into a mathematical model which represents the metabolic network of the organism. This allows us to perform simulations and optimizations by using mathematical methods in order to make *in silico* predictions of metabolic states and generate hypotheses.

Creating a GEM is a time-consuming process. No procedure is yet available for the fully automated reconstruction of GEMs and still a lot of manual curations are required. However, numerous computational methods and protocols have been proposed to make this process semi-automatic and to generate high-quality genome-scale models [64,65]. Furthermore, there are repositories of publicly available reconstructed genome-scale models for several organisms such as BIGG [66] and SEED [67] databases.

GEMs are proven to be valuable in many applications in basic biological studies, metabolic engineering, biotechnology and pharmaceutical researches for drug design. However, since they are derived from genome annotations that are themselves incomplete, they may not fully capture the metabolic and enzymatic capacity of the organisms. Consequently, several metabolic pathways remain unknown, and many reactions are still missing even in known pathways and well-studied organisms [68]. This necessitates the development of higher-level computational tools beyond those

depending on the genome sequencing and annotations, in order to fill in the metabolic knowledge gaps by *de novo* prediction and the reconstruction of metabolic pathways.

2.1.2 *In silico* pathway design

Several computational frameworks have been developed for the *in silico* pathway design in metabolic networks. The most commonly *in silico* pathway design tools offer the enumeration of pathways in two ways: (i) either they effectively combine known reactions from databases that lead to the production of a given desired compound from different organisms (heterologous pathways) [69-71] or (ii) they construct the *de novo* pathways which include not only the known reactions but also the hypothetical steps whose corresponding enzymes might not actually exist in nature [6,72-79].

The algorithms in the **first approach** are based on the graph representation of metabolic networks. Using this graph, one can search for all the possible pathways between known input and output compounds. The source graph for the pathway searching can be either limited to the metabolic network of an organism, or extend beyond a specific organism and include all available known reactions. One of the best examples of such tools is FMM (From Metabolite to Metabolite), a computational tool for the reconstruction of metabolic pathways form one source metabolite to another target metabolite among different organisms [69]. FMM and other similar methods [28,69-71,80] enable the identification and the design of new metabolic pathways based on *known reactions* previously existing in the databases. Although there are numerous achievements in the metabolic engineering research through discovering heterologous pathways using the aforementioned tools, such reconstructed metabolic pathways are based on existing metabolic maps. Therefore, such methods are limited when there is no known enzymatic step for a desired compound and one has to design a *de novo* pathway for its biosynthesis [72]. This scenario is usually the case when the target compound is a non-native compound for the known chassis organisms.

Algorithms of the **second approach** enable postulating putative metabolic pathways that are of great interest in synthetic biology. A comprehensive algorithm for the *in silico* prediction and design and further feasibility evaluations of *de novo* pathways is a significant driver for the success of metabolic engineering and various such tools have

been developed in the past decade (Table 2.1). We will elaborate more on the technical aspects of such algorithms in the next section.

Table 2-1: Available *de novo* pathway reconstruction tools and their availability

Name	**Tool development and applications**
BNICE.ch	[6,72,81-84]
DESHARKY	[75,85]
ReBiT http://www.retro-biosynthesis.com	[78]
Method developed with Cho *et al.*	[73]
RetroPath http://www.issb.genopole.fr/~faulon/retropath.php	[74,86-90]
SimPheny	[76]
GEM-Path	[77]

2.1.3 Retrobiosynthesis for *de novo* pathways design

Retrobiosynthesis, a promising approach for the *de novo* pathway design, is inspired from the retro-evolution hypothesis proposed in 1945 by Norman Horowitz [91,92] and has its origins in retrosynthetic organic chemistry. Retrosynthetic analysis starts by defining a target molecule that we are interested to produce and walks backward through the known chemical transformation rules to modify the target molecule and identify the possible precursors and reactions [93,94]. This basic concept of walking backwards from a molecule and using the biotransformation rules to reconstruct biochemical pathways is also used: (i) to find novel pathways for the biodegradation of pollutants [79,95] and (ii) in generating hypothetical pathways for the metabolites and lipids that are found in metabolomics and lipidomics studies but their metabolism is unknown [84].

In retrobiosynthesis, the goal is the production of a target molecule through the enzymatic biotransformation steps occurring in the metabolic network of

microorganisms. This analysis results in finding *de novo* pathways that connect the target molecule to either a cellular metabolite or to a biochemical feedstock by using natural or engineered enzymes.

Before a *de novo* pathway could be built in the laboratory and integrated in a microorganism, it should first be designed and evaluated. While intuition and manual design can assist in postulating novel pathways, it is not sufficient to guarantee the generation of all the potentials and the selection of the most efficient ones [55,72,73,89,96-99]. Hence, the computational prediction tools are indispensable for retrobiosynthesis analysis, not only for assisting with generating novel hypotheses, but also for screening for the most efficient pathways. Computational frameworks result on the extensive generation of all possible *de novo* biosynthetic pathways that allows for the exploration of the entire space of feasible biotransformations in a given cell [6,72-79,81].

The combinatorial explosion is the most important risk associated with these approaches, since they generate compounds and reactions which may or may not actually take place in nature. Therefore, the next crucial step is to screen the generated biosynthetic pathways through feasibility studies.

Following the increasing demand for designing *de novo* pathways in the metabolic engineering research, computational tools for the retrobiosynthetic analysis attracts more attention. These tools are becoming one of the most important steps for a reliable metabolic engineering strategy.

BNICE.ch (Biochemical Network Integrated Computational Explorer) which has been developed by Hatzimanikatis et al. in 2005 [72], is a pioneer computational framework for the *de novo pathway design and evaluation.*

One of the important modules of BNICE.ch is the retrobiosynthesis analysis. From our experience in developing the retrobiosynthesis framework of BNICE.ch and the analysis of other available tools, we propose a retrobiosynthetic workflow composed of three main steps, where each step requires certain *technical design elements* to be implemented (Figure 2.1).

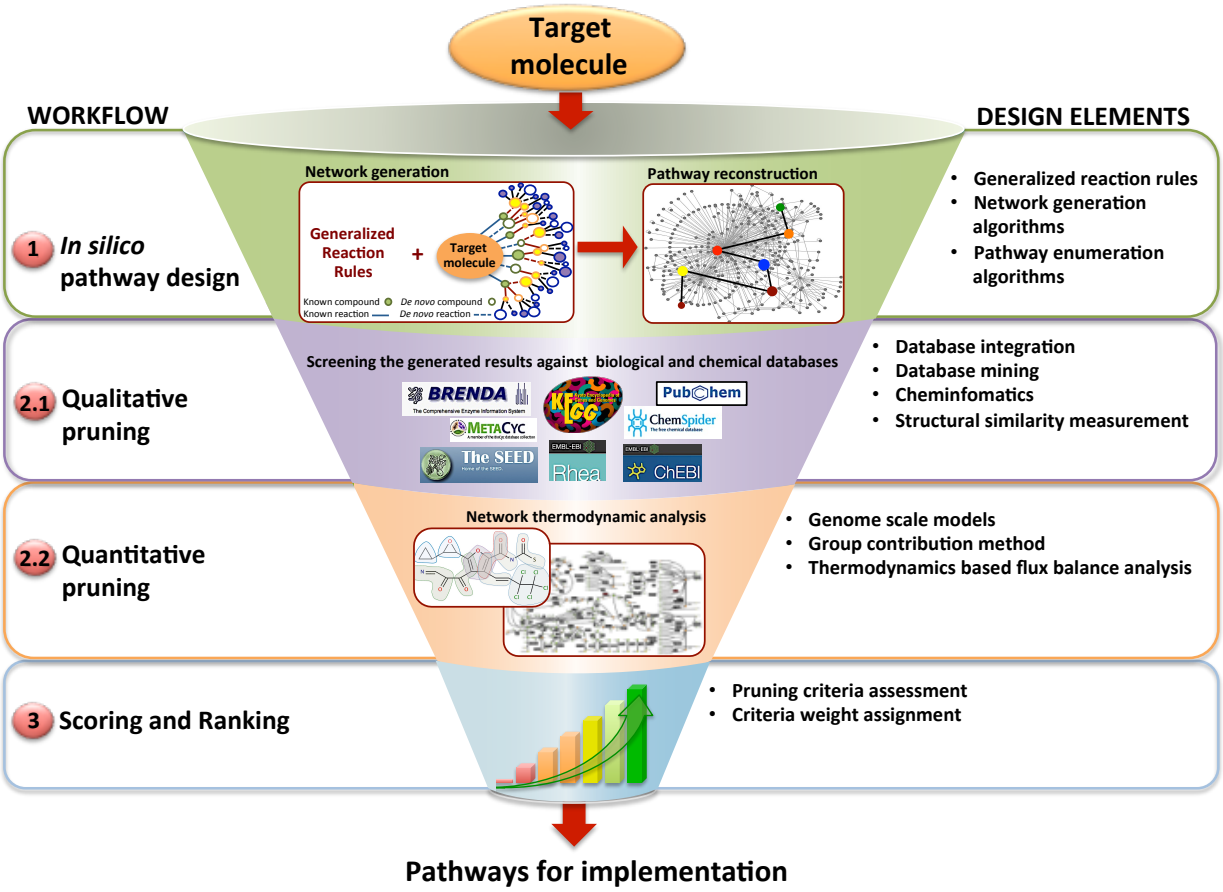


Figure 2-1: Our proposed workflow for a retrobiosynthesis framework start with the selection of a target molecule and results on top-ranked synthetic pathways for implementation in a chosen organism for the biosynthesis of the target molecule.

In the next section, we introduce BNICE.ch, and we describe further developments and extensions to this framework that have been made during this thesis. The applications of the developed methods are demonstrated in chapters 3 to 7.

2.2 BNICE.ch framework

BNICE was initially developed in 2005 for the exploration of metabolic networks [72,81]. In the course of this PhD project, several modules and methods have been integrated in BNICE.ch for enhancing its capabilities and extending its applications. Furthermore, we have applied BNICE.ch in several research and industrial projects. The achievements of these research projects will be discussed in the following chapters. Furthermore, we have applied the retrobiosynthetic module of BNICE.ch in several

industrial projects, the results of which are cataloged in a web-based database and cannot be discussed here due to confidentiality agreements. However a representative example of such analysis is available on the website (lcsb.epfl.ch/database).

In the following sections, we present BNICE.ch, an established computational framework to *design, evaluate, rank and visualize* promising *de novo* pathways for several applications ranging from metabolic engineering to drug design. Below, we list and describe the technical design elements of BNICE.ch (Figure 2.2).

2.2.1 Internal and external databases

2.2.1.1 Generalized enzymatic reaction rules

One of the key design elements of BNICE.ch is a database of "biochemical transformation rules". These rules mimic the function of enzymes and serve as *in silico* enzymatic actions. As there are a large number of characterized enzymes, one can organize those that perform similar reaction mechanisms into "*generalized enzymatic reaction rules*" [72,81]. After their introduction in BNICE.ch in 2005, the concept of generalized reaction rules has been adopted by several other similar methods [73,74,76-78].

BNICE.ch has an in-house made database of 582 manually curated generalized reaction rules that translate the biochemical knowledge of enzymatic reaction mechanisms into a generalized mathematical format. Around 90 percent of known enzyme-catalyzed reactions can be represented by a set of 291 manually curated, bidirectional reaction rules that can be considered as "*in silico* enzymes". Each of the 291 forward reaction rules comes with a reverse reaction rule catalyzing the opposite direction of the reaction. Thus, the whole set of reaction rules contains 582 forward and reverse rules.

The generalized reactions rules are developed based on the Enzyme Commission (EC) classification of known biochemical reactions in enzyme reaction databases [100]. The Enzyme Commission classification is a numerical classification scheme for enzymes. It uses a four-digit number (EC number) that represents a progressively finer classification of the enzyme. If different enzymes (for instance from different organisms) catalyze the same reaction, they are assigned the same EC number.

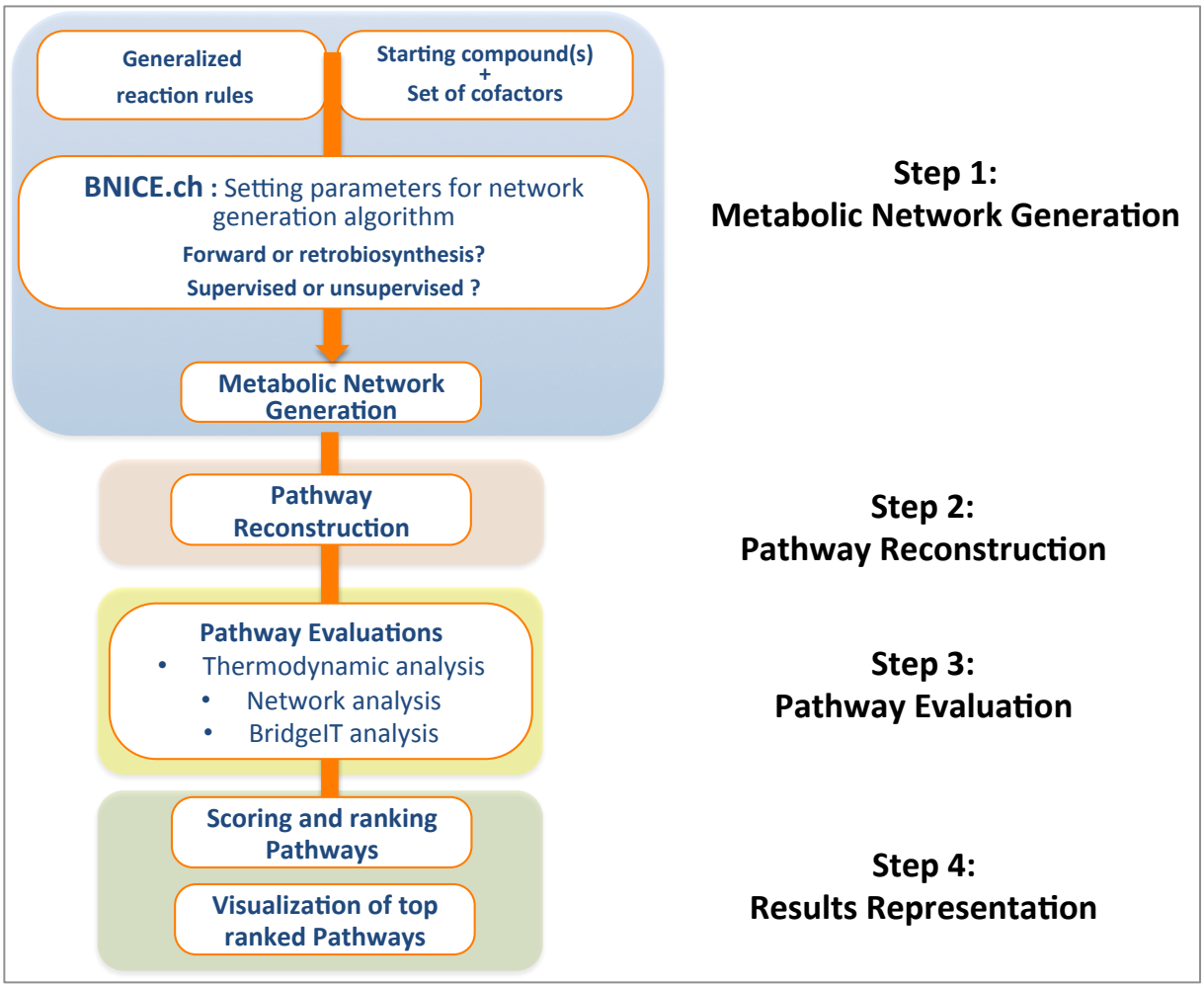


Figure 2-2: Workflow of the BNICE.ch framework is categorized into four main steps and each step may include more than one design elements.

We have developed a procedure for extracting the generalized reaction rules from known enzyme reactions. Since the generalized enzyme reaction rules are not substrate specific, we formulate a new rule based on all existing specific (4^{th} level) reactions in the corresponding 3^{rd} level class: (1) we find repeated patterns for the reactive sites of the substrates in the 4^{th} level enzymes, then (2) we represent these repeated patterns using bond-electron matrix (BEM) [101], and finally (3) we define a generalized reaction rule consistent with the structural similarity (exact match) of the reactive sites of substrates as derived by repeated patterns.

Each generalized reaction rule has three representative matrices (Figure 2.3): the **first** one is the BEM representation of the reactive site of substrates for a class of enzymes that follow a unique reaction mechanism.

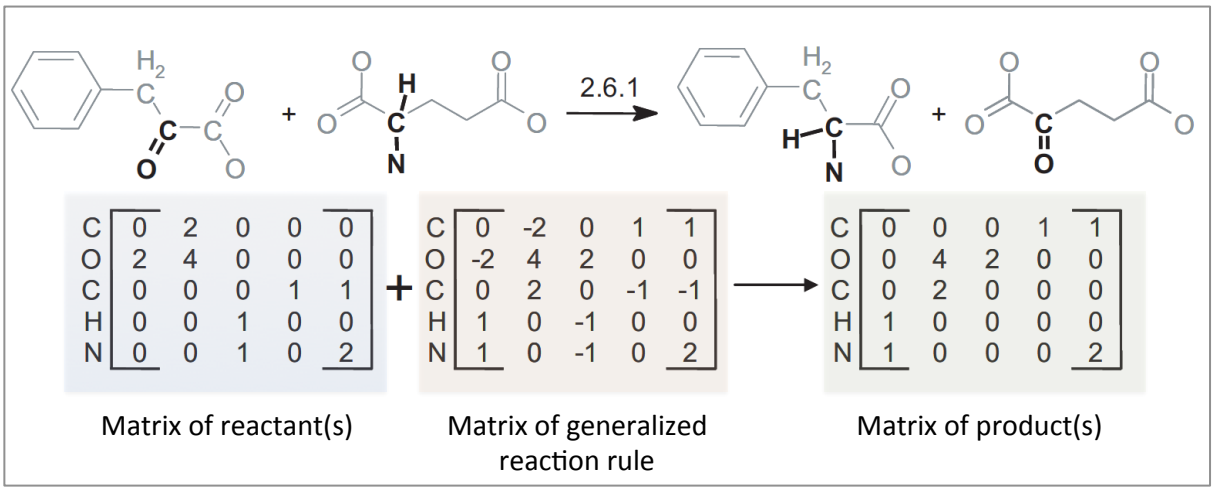


Figure 2-3: Example of the action of a generalized reaction rule on the substrate of a reaction and the formation of products. The reaction rule matrix is added to the BEM of the reactants to produce the BEM of the products.

The **second** is the reaction matrix that determines the change in the bonds during the overall reaction and is used to apply the reaction through matrix addition to the BEM of the reactant molecules. Negative numbers in the reaction matrix correspond to the cleavage of bonds and positive numbers correspond to the formation of bonds. Finally the **third** matrix is the BEM of the products that are formed when the matrix representing the enzyme catalyzed reaction is added to the BEM for the substrates (details and examples can be found in [6,72,82,95,102,103]). During my thesis, we systematically revised and expanded the generalized reaction rules. Until this work on the revision of the rules, not every rule had the corresponding reversible action and most of existent reveres rules had not been properly curated. We systematically curated also the reverse rules and now for every "generalized enzyme reaction" we have two rules, one for each direction. Prior to the work in this thesis, BNICE.ch had a set of 86 rules and during the work of this thesis this number increased to 582 that accounts for

291 rules in each direction. More details about the new set of generated reaction rules will be discussed in Chapter 3.

2.2.1.2 Biological and chemical databases integrated in BNICE.ch

A very important aspect of BNICE.ch is the information from available databases integrated into the framework, which allows us to screen our results against all known biological compounds and reactions. If a specific reaction of the database can be replicated using a generalized reaction rule, we denote the reaction as being "covered" by BNICE.ch. The percentage of reactions in a database that can be replicated with our reaction rules is called “coverage”, which is an important indicator for the performance of BNICE.ch and makes it distinguishable from other similar tools.

One has to notice that the “coverage” is a moving target; since it depends on the actual number of “known reactions” which is not fixed and is increasing over years. During this thesis, by curating several new reaction rules, the coverage of BNICE.ch increased from 48 % of KEGG 2008 (~6000 reactions) to ~90% of the KEGG 2014 (more than 9000 reactions).

The most important biological databases as sources for metabolic data are the Kyoto Encyclopedia of Genes and Genomes (KEGG) [60], SEED [67,104] , MetaCyc [61], and ChEBI [105] databases and their information is integrated in BNICE.ch. KEGG is a manually curated, reliable source of information that is used in this work as the gold standard for assessing the performance of our method as well as for the validation of reaction mechanisms through KEGG RPAIR database. KEGG RPAIR contains reaction mechanisms as well as information about atom-atom correspondence [106].

In addition to biological databases, we have integrated the enormous amount of information of PubChem database [107,108] (~2 million unique entry for chemical structures)- the biggest available database for compound structures- in BNICE.ch.

In Section 2.2.4 we elaborate more on how we match the generated compounds and reactions against the mentioned databases.

2.2.2 Metabolic network generation in BNICE.ch

BNICE employs an automated *network generation algorithm*, which works in an iterative fashion. We use “generation” and “iteration” interchangeably in this work.

Starting with a set of input molecules, the algorithm works as follows:

1) Every molecule is checked for reactivity, i.e., it is evaluated to find if it has the appropriate reactive sites (functionalities) to undergo the reactions corresponding to the specified list of reaction rules.
2) Upon acting on a molecule, the generalized reaction rules recognize the biological reactive sides of molecule and apply the biotransformation by which the atoms and bonds rearrange to form the product.
3) Next, all reactants are placed in a "reacted" list, and all products from these reactants will be placed in an "unreacted" list if they are molecules that have not been specified or generated before. This completes the first step and is defined as 'generation 1'.
4) Each molecule in the "unreacted" list will be checked for its reactivity, the reaction rules will be applied, and new "reacted" and "unreacted" lists will be created for the 'generation 2'.
5) The procedure is repeated iteratively and an iteration count is maintained as new molecules are created, keeping track of the generation number of each species, which corresponds to the number of steps required to create a given product from the original reactant(s).
6) A maximum generation number can be specified, and thus the generation number can be used to determine if a given molecule from the "unreacted" list may react in the next generation. Once the generation number reaches the specified maximum, the newly created molecules are placed in the "reacted" list, which marks the terminal point of the reaction network.

Being designed as "general", the reaction rules are capable of acting upon a wide range of substrates in addition to the specific native ones. Therefore, repeating this process iteratively results in the generation of a biochemical network of all theoretically possible compounds and reactions, including those that have no known experimental counterpart (*de novo* compounds and reactions).

Aside from the "number of iterations", the number of generated compounds and reactions and consequently the size of the generated metabolic network depends on several other adjustable parameters that are defined in BNICE.ch and will be further

discussed.

2.2.2.1 'Forward' vs. 'retrobiosynthesis' network generation modes

BNICE.ch has two different running modes, "forward" versus "retrobiosynthesis", which allow us to reconstruct a metabolic network in two different approaches based on the nature of the study. The forward algorithm explores all the possibilities between the compounds in the "unreacted" list to generate all possible reactions between them. This running mode has been applied to reconstruct several metabolic networks, starting from a set of substrates along with the generalized reaction rules. We demonstrate the results of the *forward running mode* in Chapter 4 for the reconstruction of part of the lipid metabolism and in Chapter 5 where we reconstruct the *E.coli* core metabolic network.

In the retrobiosynthetic algorithm, we take a slightly different approach based on the concept of retrosynthesis. A retrosynthetic analysis starts from a desired target compounds with the goal being to find all possible reactions steps (pathways) from this compound to potential substrates. The aim at each further iteration is to generate a reaction step that brings us closer to the *potential substrates*. We match the generated compounds in a network against the metabolites of the chosen organism and we define the potential substrate as any metabolite that is native metabolite for the organism and exits in the generated network of the target compound.

Therefore, in the retrosynthetic mode, we allow the set of metabolites in the "unreacted" list in each generation to only react with other starting compounds (a.k.a. reactive starting compounds) but not with each other. After each generation, the new set of generated compounds is allowed to react with only the set of reactive compounds of previous generations in order to generate all possible biochemical reactions. This feature would allow us to always stay connected to the target compound through linear pathways of enzymatic reactions, and in addition, it helps us to explore more generations as it reduces the number of generated compounds and reactions.

We have applied the retrobiosynthetic approach of BNICE.ch to analyze the lipid metabolism and to integrate the lipid structures from lipidomics data into metabolic pathways [84], which will be covered in Chapter 4.

Furthermore, we have successfully applied the retrobiosynthesis feature of BNICE.ch in several industrial projects for finding *de novo* pathways for the biosynthesis of chemicals with no known biosynthetic routes, an example of such analysis is provided on our website (lcsb.epfl.ch/database).

2.2.3 Pathway enumeration algorithm

We analyze in further details the identified compounds and reactions in order to determine the architecture of their synthesis and degradation pathways. We use *a "pathway reconstruction algorithm" that* constructs all possible pathways from a given substrate to the target molecule in the generated network of compounds and reactions. In general, the pathway reconstruction algorithms perform either a graph-based search in the network or use optimization-based methods to identify possible pathways from potential substrates for the synthesis of the target compound in the generated metabolic network [82,109,110]. We implemented both features in our pathway search algorithm.

2.2.4 Pruning the generated data

The *in silico* design of the *de novo* pathway risks a combinatorial explosion in two aspects. **First**, in the network generation process, the action of the generalized reaction rules on the compounds results in the generation of all possible compounds and reactions which may or may not actually take place in nature. The number of such compounds and reactions increases exponentially at every iteration of the network generation algorithm. **Second**, due to the combinatorial nature of the pathway enumeration step, an enormous number of pathways from a substrate to the same target compound are generated. Thus, the very important next step is the evaluation of proposed compounds, reactions and pathways and the selection of the most feasible enzymes, reactions and pathways to be tested in laboratory. In BNICE.ch we perform the pruning analysis by two strategies:

(1) Qualitative pruning of generated data

(2) Quantitative pruning of generated pathways

2.2.4.1 Qualitative pruning of generated results

Qualitative pruning of the generated data is the process of surveying which portion of the obtained information is already *known*, which portion is *novel,* and how similar is the novel information compared to the known data, i.e., the metabolites, reactions and pathways existing in the databases. These databases are biological, such as KEGG [60] and Metacyc [61], and chemical such as PubChem [111]and ChEBI [112]. Qualitative pruning is in general independent of the organism of choice.

By screening through existing databases, in addition to differentiating between known and novel knowledge, we also directly capture available biochemical properties for the compounds and reactions.

2.2.4.1.1 'Supervised' vs. 'unsupervised' approaches

The size of the reaction network that BNICE.ch generates depends on different constraints predefined in our algorithm. The most important factors that have a major effect on the size of the generated network of reactions are:

(i) the number of generalized reaction rules and

(ii) the allowable "*search space*" for the generated compounds and reactions.

These are both introduced, as an input parameter for the algorithm and the consideration of these two parameters would result in a "*supervised network generation*" approach as opposed to the "*unsupervised network generation*" approach.

In a supervised network generation mode, based on the question we address, we allow *specific* generalized reaction rules as opposed to all the reaction rules. An example of such approach is discussed in Chapter 5 where we investigate the biochemistry of *E.coli* core metabolic network. In this study, we only apply those reaction rules that have relevant enzymatic reactions in the core metabolic network of this organism.

In addition to the selection of the generalized reaction rules, during this thesis, we have introduced the notion of supervised network generation through the *adaptable search space* in the network generation process. The adaptable search space allows searching within a domain of metabolites and reactions that is predefined as a parameter. The supervision can be applied for the generated compounds or reactions or both leading to:

• The selection of the compound search space, i.e., at each iteration, we keep only the compounds that are part of a biological or chemical database or both (versus keeping all known and novel compounds at each iteration).

• The selection of the reaction search space, i.e., at each iteration, we allow only known reactions of KEGG or the reactions that are part of a specific database (versus keeping all known and novel reactions at each iteration).

Therefore, at each generation we retain in the 'unreacted' (product) list only the compounds and reactions that exist in the indicated reference databases.

By introducing these three types of constraints (for the selection of the reaction rules, compounds and reactions), we can create different project-specific supervision modes. The level of supervision is defined for each of the three input constraints, and the different levels range from all possible reactions and compounds (known and novel) down to pathway-specific compounds and reactions only.

In the case of *unsupervised network generation*, we apply all generalized reaction rules for the network generation and do not screen the results until the end of the job where we crosscheck against databases to differentiate the known and novel knowledge.

Figure 2.4 is an example of three different BNICE.ch studies on glucose that demonstrates how the choice of the constraints would dramatically affect the size of the network.

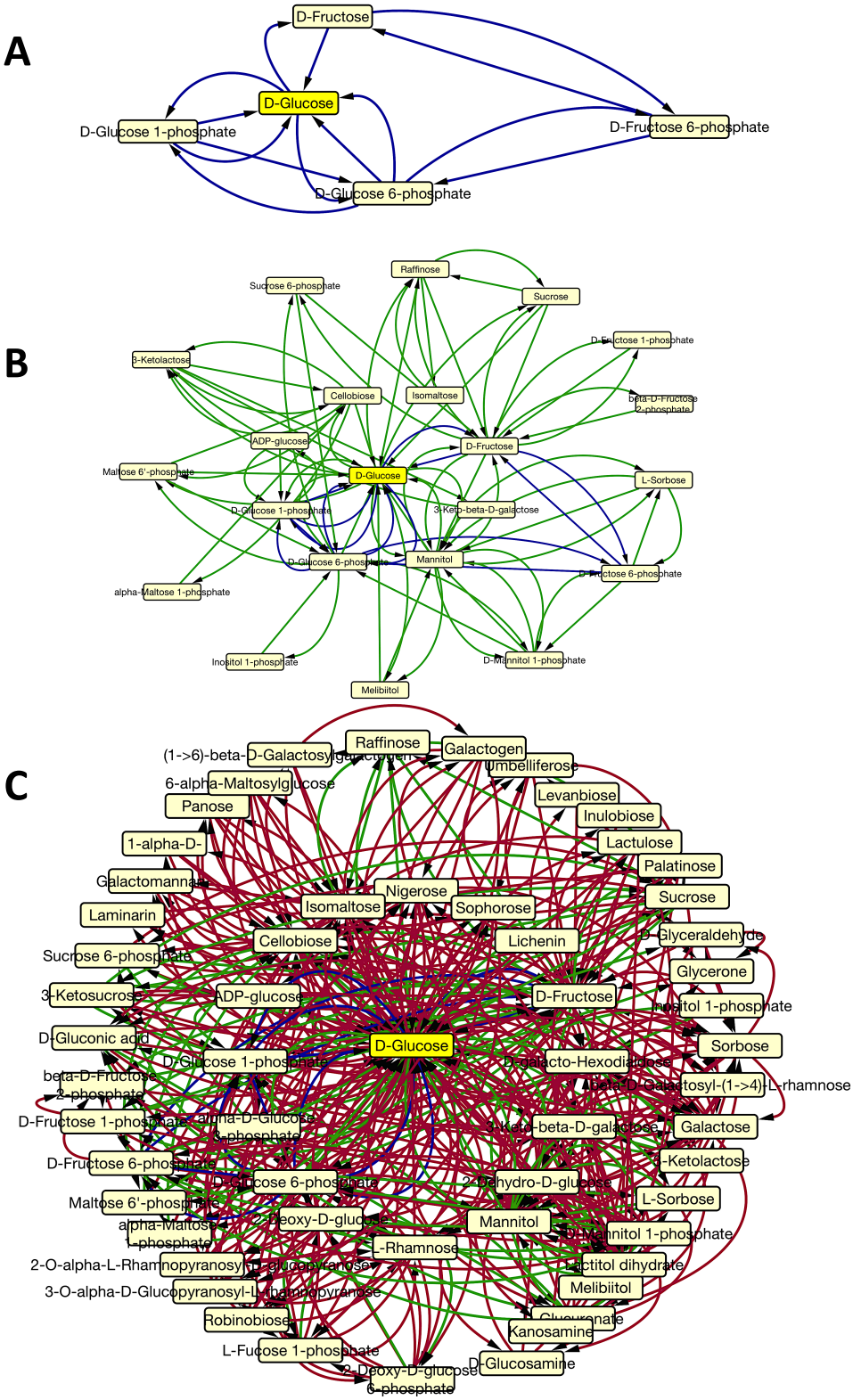


Figure 2-4: Forward network generation with BNICE.ch at three different supervision modes to investigate the impact of the supervision level on the size of the generated network. The level of supervision for the generalized reaction rules and compounds are the same in the three cases, the

level of supervision of reactions changes from only *E. coli* reactions in **A**, to KEGG reactions in **B** and finally to unsupervised reaction generation in **C**. The inner circle shows the results of the first iteration and the outer circle shows the generated compounds in the second iteration. Green connections represent KEGG reactions, red are novel reactions and blue are *E. coli* reactions.

In all the three examples, we used glucose as the substrate along with the set of 12 cofactors (NADH, NAD^+, H^+, Bicarbonate, CO_2, O_2, Water, ATP, ADP, Phosphate, Acyl-CoA and ammonia) and the whole set of 582 generalized reaction rules. We ran the algorithm for two iterations. The supervision level for the generalized reaction rules and compounds is the same for the three examples.

The smallest network (2.4 A) shows the results of the reaction supervision level to only *E. coli* reactions (5 compounds and 14 reactions).

Figure 2.4 B shows all KEGG reactions that are generated by BNICE.ch in two generations and contains 20 compounds and 77 reactions. The network size increases if we allow all the KEGG reactions (2.4 B). In the last example (2.4 C), all possible BNICE.ch reactions, starting from glucose and between the known KEGG compounds are explored. This reaction network includes 161 compounds and 808 reactions and shows the significant increase in the size of the network when we do not constraint the reaction level.

Table 2.2 summarizes different *levels of supervision* with their corresponding actual statistics. For instance, if we want to work with a network of the *E. coli* core metabolism, we would run BNICE.ch on the supervision level "Network", meaning that we only include specific reaction rules (i.e., 45) and we allow the native *E. coli* compounds (i.e., 67) and reactions (i.e., 76) to be produced. Such constrained level of supervision is especially important for the atom-level reconstruction of metabolic network that will be discussed in Chapter 6.

Depending on the application, we may want to adapt the levels of supervision independently. These features allow the efficient arrangement of the results based on the knowledge that exists in databases and address the risk of combinatorial explosion.

Table 2-2: We defined 5 levels of supervision in BNICE.ch. The supervision level can be chosen for each input constraint .The darker the color of the cell, the smaller the number of generalized reaction rules and the search space for the generated compounds and reactions. The less constrained possible case is to consider the search space of a database such as KEGG for the

reconstructed metabolic networks. In this case, all the generated compounds and reactions will be checked against KEGG, and if they exist in KEGG, we keep them with the corresponding KEGG identification numbers (Ids). The most restricted constraint is to limit the choice of the generalized reaction rules and the search space to a pathway, e.g. Glycolysis, and to only keep the generated compounds and reactions that belong to this pathway at each iteration. We used the KEGG 2014 for the database level of supervision and the information from the genome scale reconstruction of *E. coli*, *iJO1366* [113] for the organism level. In the chapters 3 to 7, for all the applications of BNICE.ch, we present this table and we highlight the level of supervision that is used in each study.

Supervision level		Input constraint		
	Example	**Reaction rules**	**Compounds**	**Reactions**
Unsupervised	-	291*2	Unlimited	Unlimited
Database	KEGG	291*2	17,343	9972
Organism	*E. coli*	150*2	1'039	1'387
Network	Core metabolism	45*2	67	76
Pathway	Glycolysis	10	12	16

The qualitative pruning can be also applied in the pathway enumeration step. During this thesis we have also implemented the notion of *supervised pathway enumeration* that allows enumerating "a set of viable pathways" rather than all possible pathways. Applying this new feature, we evaluate pathways based on the knowledge of compounds and reactions in databases. For instance, we can enumerate only the pathways with a pre-specified percentage of their steps existing in biological databases as known enzymatic reactions.

2.2.4.2 Quantitative pruning of generated pathways

Once we reconstruct a metabolic network of compounds and reactions and further enumerate *de novo* pathways of interest and screen them against databases, the next step is to perform the feasibility analysis. Such analysis is performed in order to determine the suitability and the performance of individual pathways and to quantitatively prune the proposed pathways to a set of most biologically feasible ones. Quantitative pruning is generally context dependent on the chassis organism. Different

metrics can be applied to evaluate the likelihood of *in silico*-designed pathways to be proficiently implemented in an organism.

One very important metric is the thermodynamics of the reaction steps and consequently the synthetic pathway which allows us to discard the pathways that are energetically unfavorable in the first place. To do such thermodynamics analysis, we used a Group Contribution Method to estimate the Gibbs free energy for metabolites and consequently the reactions [114]. This method has been developed in Hatzimanikatis lab and is used in several other computational frameworks for estimating the thermodynamics feasibility of the synthetic pathways [73,76,77]. Furthermore, in BNICE.ch, we apply constraint based modeling by incorporating the synthetic pathways one at the time into the genome scale model of the chosen organism and performing Thermodynamics based Flux Balance Analysis (TFBA) [47,115]. This further step allows us to adjust the estimated Gibbs free energy with respect to the metabolite concentration, ionic strength and pH to get closer to *in vivo* conditions. By performing a TFBA analysis, we guarantee that the obtained pathways are feasible with respect to mass balance (stoichiometrically), we assess the network thermodynamic feasibility of generated pathways, and we thus quantify their overall effects on the metabolic profile of the organism by calculating the energetic cost and the change in the biomass yield for each molecule of the generated product [116,117]. Furthermore, one of the most important outcomes of TFBA for biotechnological applications is the pruning and ranking of pathways based on the maximum production yield of the target molecule from each individual synthetic pathway. In Chapter 5 we elaborate more on the importance of TFBA analysis as a further step to quantitatively prune the wealth of information generated with BNICE.ch.

2.2.5 BridgIt analysis

Another compelling aspect of the interactive analysis with databases is the structural similarity comparison of the substrates and products of the generated *de novo* reactions to the substrates and products of the known reactions. We can quantify the results of such a comparison by using different chemoinformatics metrics such as ”compounds

fingerprints comparison" using "Tanimoto distance" [118], and assign to novel reactions a similarity score with respect to the existing reactions.

Using such a metric, one can propose potential enzymes for the *de novo* steps of a pathway based on their structural similarities to the known reactions. Through this reaction structural similarity assessment, we can further assign to novel reactions gene and protein sequences that could be used in evolutionary protein engineering and computational protein design for the experimental implementation of the novel pathways.

BridgIt has been recently developed in our lab as a complementary tool to BNICE.ch for assessing the structural similarity of the reactions. BridgIt is based on the hypothesis that chemically similar reactions share similar sequences, based on the "lock and key" principle, which is also used in the protein docking methods. The enzyme is considered as the "lock" and the ligand (molecule) as a "key". If a molecule has the same reactive sites as the native substrate for a given enzyme, and similar surrounding structure, it is expected that the enzyme would catalyze or could evolve to catalyze the same transformation on this molecule. Based on this hypothesis, if two metabolic reactions have the same reactive site and similar surrounding atoms and bonds around the reactive site, it is highly plausible that these metabolic reactions are catalyzed by the same enzyme(s), or the corresponding enzymes share the same E.C. classification (up to the third class). Also, it is highly probable that the genes functions responsible for these reactions (biotransformation) share sequence similarity.

The initial idea behind BridgIt development is to use the reactions' structural similarities as a measure for assigning genes to the hypothetical reactions generated in BNICE.ch. This method is further applied for finding and assigning protein sequence for the orphan reactions. Orphan reactions are enzymatic reactions that lack an associated protein sequence [119]. Due to the lack of reliable annotations and the drawbacks of homology-based predictions, a large part of known enzyme activities is still missing an associated gene sequence. Since BridgIt is not functioning based on the sequence similarity, and instead, takes into account the structural similarity of the reaction, it is a promising tool to complement the existing methods for assigning protein sequences to orphan reactions.

BridgIt translates the structural definition of a reaction to a mathematical form, so-called a vector, and compares these vectors using "Tanimoto distance". It has an integrated reference database which accounts for all KEGG non-orphan reactions.
It compares any given reaction (for instance an orphan reaction) with all the reactions in the reference database and assigns a Tanimoto similarity score for all the comparisons it does. The Tanimoto score indicates how similar the given reaction is to each of the reference reactions. It varies between 0 and 1, where 1 is considered as a high similarity and 0 indicates no similarity.

2.2.5.1 New features of BridgIt

During this thesis, we applied BridgIt in several projects for the further analysis of the hypothetical reactions generated in BNICE.ch. If we consider the *de novo* reactions generated with BNICE.ch as "*theoretical orphan reactions"*, by applying BridgIt we can find the most structurally similar known reaction along with its similarity score to the *de novo* reaction. Consequently, we can propose gene sequences for their further practical implementation in metabolic engineering studies.
In the course of my thesis, we made comprehensive case studies using BridgIt. After the analysis of the generated similarity scores, we came up with two different strategies to be incorporated in the BridgIt algorithm to improve its predictions. The first method is applicable for both orphan and theoretical orphan reactions, and the second approach is oriented for the novel reactions (theoretical orphan) that are generated with BNICE.ch and are associated with a third-level EC number.
In the **first method** we first exclude the cofactors from the generated reaction vectors to make a cofactor-free reference database. BridgIt initially considered all the compounds in the reactions as "substrates", and in the reactions in which pairs of cofactors with big molecular structures participate, for instance NAD/NAD^+, the results were not accurate. Consider two reactions that both have NAD/NAD^+ in their mechanism: they would be scored as "very similar" since they share a big portion of their vector which is NAD/NAD^+, and therefore the biotransformation that happens in the course of the reaction usually becomes negligible.

Therefore, we extracted from databases a list of molecules that are frequently considered as biological cofactors, and we incorporated a new feature to BridgIt, in order to exclude these cofactors from the reactions before constructing the reaction vectors. Therefore, the new feature of BridgIt compares only the biotransformations that happen in the course of the biochemical reaction.

We incorporated the **second feature** for the further analysis of the novel reactions generated by BNICE.ch in order to obtain a more precise similarity score, by employing the wealth of information provided for the *de novo* generated reactions that is a third-level EC number.

Our new algorithm first looks at the known digits of the EC number of a novel reaction. It then extracts the reactions in the reference database that share a similar third-level EC and compares the novel reactions only with them.

It is also possible to set a predefined parameter to compare the novel reaction with all the known reactions that share the *same second level EC*, or with those reactions that share the same *first level EC*. One of the advantages of this new feature is its time efficiency.

The comparison that BridgIt does for each reaction against all stored reactions in the reference database is intensively time consuming, and with BNICE.ch project that generates thousands of reactions, it appears as a limitation. With the new method (second feature), since fewer comparisons have to be performed, BridgIt will run much faster than with the original method. The new method excludes false positive comparisons by filtering out the reference reactions that have no similarity with the novel reaction due to the different enzyme class that they belong to.

We have also investigated the performance of BridgIt, when we introduce these two features simultaneously. This new feature first filters the cofactors from the reactions and makes the comparison of the novel reactions with only the reactions that have similar corresponding EC classification.

In Chapter 3, we perform BridgIt analysis for a large list of novel reactions generated by BNICE.ch and we demonstrate how incorporating these two features improved BridgIt predictions.

Ultimately, the result of BridgIt is a proposed known reaction along with a similarity score that can be used as scoring and ranking criteria for the evaluation of *de novo* pathways.

2.2.6 Scoring and ranking the biosynthetic pathways

Reconciling the metrics obtained in the qualitative and quantitative pruning strategies together with the results of BridgIt analysis, one can define a scoring and ranking feature which combines and scales different factors and assigns an overall score for the prioritization of *in silico* generated pathways. Using such a score, the collection of generated pathways can be filtered on the basis of biochemical knowledge and available experimental data.

For instance, in a retrobiosynthetic approach that the goal is to find promising pathways for the production of a desired chemical, such a score gives the capability to pinpoint the best candidate synthetic pathways that are most likely to produce the desired target molecule and that can be implemented in the metabolic network of the chassis organism.

In BNICE.ch we rank the pathways based on the following individual scoring metrics (max score of 1 per criterion) to determine the overall pathway score (full score = 5):

I. Thermodynamic feasibility: 1 for pathways that are thermodynamically feasible and 0 otherwise (a pathway is thermodynamically feasible if all reactions in the overall pathway is thermodynamically feasible in the direction that maximizes product formation, we obtained this score by performing TFBA analysis).
II. Pathway length score = (1 / number of reactions in pathway).
For practical reasons, when it comes to the implementation of *de novo* pathways, shorter pathways are preferred to longer ones since fewer steps need to be engineered and protein costs would be minimized.
III. KEGG reaction score = (number of known (KEGG) reactions in pathway)/ (length of the pathway). The fewer novel steps, the less enzyme engineering for the pathway.
IV. Network feasibility analysis: maximum product yield.

By embedding the novel pathways into the genome scale model of the organism, we can investigate the effect of the pathway on the original network. In the case of the retrobiosynthetic analysis, we can calculate the maximum yield for the production of the desired compound.

V. Reaction similarity score (BridgIT score) for novel reactions, in order to evaluate the likelihood of the hypothetical generated reactions.

Instead of calculating the overall score, one can also do the ranking of the score of a certain criterion as the primary ranking, and then perform a secondary ranking based on another criterion, and so forth. For instance, one approach can be choosing the pathways with maximum (or economically feasible) yield, and among them the ones with the minimum number of novel reactions as their implementation will involve a smaller number of engineering enzyme steps.

One should be careful when applying certain criteria used for pruning the obtained data that it is a multi-objective problem and different applications might give different weights to different criteria. Moreover, some of these criteria depend on the current technologies and although some of the pathways can be ruled currently as infeasible, new technologies can enable their realization in the future.

2.3 *In silico* atom mapped network integrated computational explorer

So far we discussed different computational methods for the analysis of the metabolic reactions at the "metabolite" level, since the mathematical (graph) representations of metabolic networks are traditionally done on the metabolite level [27]. Nevertheless, knowledge about the atom transition in the metabolic network is crucial for elucidating the mechanism of enzymatic reactions and calculating the reaction fluxes. During this thesis, we developed a computational method to introduce an additional level of detail by considering the track of individual atoms through metabolic reactions [27].

Several algorithms have been developed to address the automated atom-mapping problem. As comprehensively reviewed in [120], these algorithms can be classified into two main classes based on: (i) finding the maximum common substructures (MCS)

between substrates and products of the reactions (common subtracted-based methods) [25,121-126]; and (ii) optimization methods [30,126-133].

Both approaches rely on the graph representation of reactions along with the graph-based pattern recognition. In the optimization-based methods that were extensively practiced recently, one has to define a relevant objective function to be optimized for the atom-mapping solution. The most common objective functions in automated atom mappings include:

i. Maximizing the size of common subgraphs between substrates and products [25,122].
ii. Minimizing the graph edge edit distance[129].
iii. Minimizing the number of bonds broken and formed [30,125,128,130,131,133].

However, even if we assume that the aforementioned algorithms do find an optimal solution, there is no guaranty that all enzymatic reactions follow the optimal way for relocating atoms from substrates to products and that the objective functions lead to biochemically relevant atom mappings. Furthermore, most of these algorithms result in a big dataset of mapped reactions since they output multiple possible atom mappings for each reaction. Therefore, manual evaluation is required for accounting the reaction mechanism in case there is more than one atom map reported. Besides, the existing algorithms provide atom mappings for a single reaction and it would be difficult to extend these mappings for every pathway and in every metabolic network of a given organism without extensive manual work.

To the best of our knowledge, there is no published algorithm for the automatic reconstruction of the *atom-mapped metabolic network* without constraints on the size of the network. Such an algorithm would be very useful in many studies, such as drug design, where knowing the fate of each atom of the candidate drug through the transformation pathways in human body would help to understand the mechanism of drug metabolism. The ability to trace the fate of individual atoms through the metabolic pathways is also useful in many applications of metabolic engineering [125,131,134] such as for the identification and the engineering of novel biosynthetic routes for the microbial production of desired compounds [28,133,135].

During the course of this PhD thesis, we developed the "iAM.NICE" (*in silico* Atom Mapped Network Integrated Computational Explorer) framework for addressing the atom-mapping problem. Our method has the capability to create *in silico* labelled metabolites as substrates, and to transfer the labels to the products according to known reaction mechanisms.

We implemented the concept of formulating generalized enzymatic reaction rules developed in BNICE.ch [72,81] for transferring the *in silico* labelled atom in a substrate to a specific position of one of the reaction products. In "iAM.NICE", we automatically generate *in silico* labelled substrates and we apply the enzymatic rules on the labelled substrates, which results in the generation of *in silico* labelled products (atom-mapped reaction). KEGG RPAIR [136] is a manually curated database that contains knowledge on atom mapping stored in the "reactant pairs" to elucidate the reaction mechanism. In KEGG RPAIR, reactions are decomposed into reactant pairs (RPAIRs) which formulate a pairwise association between one or several groups of atoms from substrate to one or several groups of atoms in products, by taking into account the contribution of these atoms (main reactant, cofactor, etc.) in the reaction. Most of the automated atom mapping methods cross checked their atom mapping predictions with KEGG RPAIR to evaluate the performance and accuracy of their algorithm [26,122,125,129-132,137].

Interestingly, in our framework, the wealth of information of KEGG REACTION along with KEGG RPAIR is implemented for formulating the generalized reaction rules. Therefore, the performance of "iAM.NICE" for predicting the atom-atom correspondences in the biological reactions and consequently metabolic networks is intrinsically validated.

We further define an *atom-mapped metabolic network* as a network that contains atom correspondence for each single reaction and for the network as a whole, meaning that an atom in the initial substrate and in its final product carries the same label.

In Chapter 6 we describe details of our methodology and we discuss the important steps "iAM.NICE" takes to:

- Automatically map atoms from substrate to product in a single reaction
- Connect these atom-mapped reactions into atom-mapped metabolic pathways
- Integrate the atom-mapped pathway in metabolic networks

Moreover, we provide the atom-mapped representation of all the reactions in KEGG database as well as an atom-level representation of the core *E-coli* metabolic network. Our results can be used for a large range of applications, starting from the identification of new metabolic routes for the microbial production of desired compounds to the design of the optimum labelled patterns of substrates, which can be a great benefit for the simulation of tracer experiments.

2.4 Biothermochemical computations for studies of metabolism

Thermodynamics plays a critical role in studying the metabolic networks. Particularly in the field of *de novo* pathway design, it is necessary to evaluate the generated *de novo* data on the basis of thermodynamic feasibilities. In this section we review the exiting computational methods for estimating the thermodynamics properties and we introduce our novel approach for obtaining such estimations.

As mentioned in Chapter 1, the group contribution method (GCM) plays a central role in the thermodynamics study of the metabolism. However, being based on experimental data, this method faces limitations in two respects. First of all, experimental data are limited for biological compounds, especially for heavy metabolites such as aminoacids and oligopeptides, saccharides, nucleosides and their derivatives. Second, since the experimental data are obtained on ambient conditions, the GCM calculations and estimations based on these data hold true for the ambient condition as well. As described in Chapter1, for many applications and especially the biological applications, it is important to calculate the thermodynamic properties at biologically relevant conditions of temperature and pressure, for example at T=37-40 for the studies of human metabolism under healthy and disease conditions.

To overcome the limitations of existing methods, one needs to improve the GCM method for obtaining predictions beyond the metabolites that have available experimental data and also to adjust the estimated thermodynamic properties to any temperature and pressure based on the nature of the study. To address both issues, we designed a pipeline for the thermodynamics analysis of metabolism, which combines GCM and *quantum chemical calculations* to compute the thermodynamics properties for a wide range of biological compounds in adjustable conditions of temperature and pressure.

2.4.1 Overview of high-level quantum chemical calculations

There are fundamental reasons for the increased use of quantum chemical calculations among chemists [138]. The theories underlying calculations have now evolved to a stage where a variety of thermodynamic properties can be predicted.

However, there is no available *ab initio* method for the estimation of thermodynamic properties at *aqueous phase* that corresponds to the biological systems.

Current advances in quantum chemical calculations and the unprecedented and ever increasing computational speed of computers has made the quantum chemical methods the most reliable approach for calculation of thermochemical quantities in the *ideal gas state*. In addition, there are already significant advances for the calculation/prediction of the corresponding *hydration* free energies and enthalpies in quantum calculation methods [139-141]. The latter feature would allow bringing the generated data in the gas phase to aqueous phase and eventually it would allow studying the thermodynamics of biological systems. Therefore, the *first step* is to estimate the thermodynamic properties at the *gas phase*.

In a recent thorough review [142], the state-of-the-art in quantum thermochemical calculations is compared with the corresponding group-contribution method (GCM) approach. Under certain conditions, the level of accuracy of current quantum thermochemical predictions in the ideal gas state compares with or overpasses the thermochemical accuracy of 1 kcal/mol [143,144] for small to moderate sized molecules with 2 to 10 non-hydrogen atoms.

The Gaussian – n (Gn) family of quantum chemical procedures [144,145] achieves the above level of accuracy for moderate sized molecules but the accuracy decreases sharply for heavier molecules. This holds true for the predictions of absolute thermochemical quantities via the atomization energy differences.

Under certain conditions, accurate predictions can be made through designing appropriate *isodesmic reactions* even without using high levels of theory and computation [143,144]. In isodesmic reactions, bond types and groups are kept the same on the two reaction sides; therefore, any flaws in theory and systematic errors will be mutually compensated. This makes the isodesmic reaction approach a

computationally efficient way to calculate the enthalpies and Gibbs free energies of formation for relatively large molecules with tens or more heavy atoms [145,146].

The key prerequisite in an isodesmic reaction approach is the availability of accurate thermochemical information on all other reactants and products beside the studied molecule. A lot of work has been done in this respect for the development of extensive thermochemical databases [147-149].

Metabolic reactions take place in aqueous environments where metabolites often exist as ions. Significant advances have been made in the field of gas-phase ion thermochemistry, both experimentally and computationally [150-157]. In a recent review[158], standard values for gas-phase basicities and proton affinities were recommended for the 20 (protein) aminoacids. Gas phase acidities for these aminoacids were also reported in recent studies [157,159]. Yet, heats and free energies of formation of ionic forms are rarely reported.

There have also been some studies on the gas-phase quantum thermochemistry of saccharides and nucleosides, but not as extensive as for aminoacids. Reliable experimental thermochemical data for these classes of metabolites, especially nucleosides, are rare.

2.4.2 Computational tools used for high-level quantum calculations

Hereby we describe the computational tools we used for the calculation of heat and Gibbs free energy of aforementioned heavy metabolites in the gas phase. The heavy metabolites of interest in this work are fairly flexible and exist in a vast conformational/isomerization space, which makes their detailed theoretical computations a challenging task [160].

For the conformer search, we used the COSMOconfX suite (Cosmologic GmbH, Germany) and the Conformer Analysis application of Spartan 14 suite (Wavefunction, USA) – a Monte Carlo / Molecular Mechanics algorithm. The first gross selection of the prevailing conformers was screened further down to a few conformers by performing energy calculations with progressively increasing basis set. We reinserted some conformers with extensive intramolecular hydrogen bonding, which were rejected by the above search algorithm into the pool for further calculations at a higher level.

Whenever available, we inserted optimal geometries from literature into the pool at this stage. The prevailing conformers were subject to further geometry optimization at progressively higher levels until the most stable conformer was identified. We performed further calculations for the isodesmic reactions with the most stable conformer or with the few (less than five) prevailing conformers.

We did Quantum chemical calculations at the DFT-D3 level with Grimme's dispersion correction (D3 London dispersion correction) with Becke–Johnson damping [161] as implemented in TURBOMOLE suite[162] with the resolution of the identity RI-J approximation[163]. Geometry optimization and vibrational frequency analysis were done with the Becke-3-Lee-Yang-Parr (B-3LYP) 3- parameter hybrid functional with Becke's popular nonlocal exchange functional and Lee/Yang/Parr nonlocal correlation functional [164,165] with the def2-TZVP (Karlsruhe segmented contracted triple-z valence quality plus polarization) basis set [162]. All geometry-optimization / frequency calculations were performed with TURBOMOLE v. 6.5 suites (Cosmologic GmbH, Germany). Using this geometry, we conducted single point energy (SPE) calculations at a higher level in two alternative ways: First, the generalized gradient approximation (GGA) B97-D density functional [166,167] was used with the quadruple z-valence quality def2-QZVPD basis set as implemented in TURBOMOLE suite with the above Grimme's D3 dispersion correction. We did the SPE calculations for all compounds considered, neutral or ionic. Second, we used the Chai and Head-Gordon ωB97X-D long-range corrected hybrid density functional [168] with the 6-311++G(2df,2p) basis set, as implemented in Gaussian 09 and in Spartan 14 suites of programs. Calculations at the second level were done on selected metabolites.

We performed vibrational frequency calculations in order to verify that the structures were minima and, also, to obtain the zero point vibrational energy (ZPE) and the thermal corrections to the enthalpy, H_v, and free energy, G_v. We calculated the latter quantity from the entropy change through the classical equation:

$$\Delta_f G = \Delta_f H - T\Delta_f S$$

The entropy term in the heavy metabolites is dominated by their many low frequencies [169], which are usually poorly approximated in the quantum thermochemical

calculations, therefore the “normal-mode” approximation may not be valid for heavy and flexible molecules.

More details of the methods we applied and the isoseismic reactions we design to calculate the thermodynamics properties for the heavy metabolites along with our results will be comprehensively discussed in Chapter 7.

Chapter 3

ATLAS of BIOCHEMISTRY

How the known biochemistry evolves if we apply the generalized enzyme reaction rules against all the currently known compounds?

SUMMARY

In this chapter we introduce the most comprehensive database of all the biochemically plausible reactions that can be generated using BNICE.ch based on the known biochemistry introduced in the KEGG database. This extension of KEGG reaction database includes ~128000 reactions that can connect two or more KEGG metabolites, Approximately 5300 reactions out of ~128000 are the KEGG reactions whereas the rest are *hypothetical novel reactions* that have never been reported to occur in nature. We applied the BridgIt method, an extension to BNICE.ch, to evaluate the structural similarity of the hypothetical proposed reactions to the known KEGG reactions.

Furthermore, to validate the consistency of our results with the known biochemistry, we compared the 2 versions of KEGG database, KEGG2012, and KEGG2014. Interestingly, 81 *novel* reactions that BNICE.ch discovered based on the KEGG2012 database appeared as *known* reactions in the 2014 version. This finding validates the consistency of the BNICE.ch *generalized reaction rules* with the enzymatic reaction rules that nature follows.

3.1 Introduction

Our knowledge about the metabolism, even for the most studied model organisms such as *E. coli* and *S. cerevisiae* is not complete, and none of the available metabolic network reconstructions for any organism is considered complete and without knowledge gaps [170]. The Kyoto Encyclopedia of Genes and Genomes (KEGG) database [60], known as the most complete source of metabolic data, increases its size and number of enzymatic reactions in each yearly updated release, which demonstrates that numerous not yet characterized enzymatic reactions are still to be found in nature.

To the best of our knowledge, there is no available database that accounts for *all the theoretically possible enzymatic reactions*, which are connecting known compounds and that are based on the known biochemistry. Such information would be of great interest not only for filling the knowledge gaps in metabolic networks, but also for metabolic engineering studies where the discovery of *novel* reactions is an important mission. The question that arises is: how the known biochemistry evolves if we apply the generalized enzyme reaction rules against all the currently known compounds?

We address this question by applying the concept of *generalized reaction rules* within the computational framework BNICE.ch. Since the BNICE.ch reaction rules are based on the Enzymatic Commission (EC) classification system, every novel reaction generated with BNICE.ch was associated to a third-level EC number, which defines a biochemically relevant reaction mechanism for novel reactions. With BNICE.ch we reconstructed ~90% of known reproducible KEGG reactions, and discovered more than 123'000 *de novo* reactions involving two or more known KEGG compounds. We organized these results in a database, named "*ATLAS of biochemistry*" that comprises all the hypothetically possible enzymatic reactions between any two KEGG compounds, or more if required by the reaction mechanism.

Since *de novo* reactions are theoretical orphan reactions, we applied BridgIt (Chapter 2) for each of these reactions to find candidate genetic sequences that code for an enzyme that is potentially capable to catalyze the novel reaction. We applied different running modes of BridgIt and we extensively compared the results of different approaches.

We analyzed our results with respect to two different versions of the KEGG database, KEGG 2012 and KEGG 2014 and we investigated if any of our *predicted hypothetical*

reactions in KEGG 2012 became known in KEGG 2014. Remarkably, we found 81 novel reactions that BNICE.ch predicts based on KEGG 2012 are reported as known enzymatic reactions in KEGG. This finding validated the consistency of our generalized reaction rules with the acknowledged biochemistry, and demonstrated that our proposed hypothetical enzymatic reactions are an important complement of the known cataloged biochemistry.

3.2 Methods

The modules and methodologies of BNICE.ch have been extensively discussed in Chapter 2. As described previously, BNICE.ch performance depends on a set of parameters that we predefine for each study. Here we list the input parameters that we used in this Chapter, and we also describe the procedural steps we took to build the "Atlas of Biochemistry".

3.2.1 Preprocessing of KEGG compounds and reactions

The KEGG database keeps growing in size: In 1999, it accounted for 5'207 reactions and 5'645 compounds. In 2004, there were 5'799 reactions and 10'739 compounds. In December 2014, KEGG stored 9'972 reactions and 17'343 chemical compounds, and both categories keep growing with every new release. The continuous discovery of *new biochemistry* indicates that our knowledge about metabolism is far from being complete, and that there is a remarkable potential for the discovery of new metabolic functionalities.

For a reaction to be reconstructed in BNICE.ch we require a *defined molecular structure* for each of involved compounds. A structural definition of compounds is needed to create the corresponding bond electron matrices (BEM) as discussed in Chapter 2. On the other hand, polymers being a repetition of monomers that already exist in the database do not add any useful information to our study. Therefore, we preprocessed the KEGG compounds and excluded from our input data all compounds that:

i. Do not have a structural definition
ii. Describe the polymer structures

Similarly, we parsed all the KEGG reactions and excluded:

i. Reactions with uncompleted molecular description of the reactants and products
ii. Reactions that involve polymers
iii. Reactions that are stoichiometric unbalanced

After we removed the compounds and reactions that did not meet the above mentioned criteria, we ended up with a parsed KEGG database that consists of 14'549 compounds and 5914 reactions.

3.2.2 Reconstructing KEGG reactions and predicting *de novo* reactions

In this section, we present a pipeline for reconstructing the *latest coverage of KEGG reaction database*. We used the following two approaches to perform the reconstruction:

In the **first approach**, we applied *all the generalized reaction rules* one at a time, to *all KEGG compounds*. We used the supervision level that allowed all possible reactions (known and novel) but only KEGG compounds (Table 3.1). The procedure was as follows:

1. Apply a reaction rule to all KEGG compounds and for only one iteration.
2. Screen the output against the KEGG database and save the results
3. Switch to the next rule and go to Step 1.
4. Stop when all 582 reaction rules are covered.

As a result of this procedure we got two lists of reactions: KEGG covered (reconstructed) reactions and *de novo* reactions.

In the **second approach**, we identified those KEGG reactions that were not reconstructed after 1st iteration of BNICE.ch, and we performed further analysis with these reactions (~12 % of the reconstructable KEGG reactions). In our pipeline, we read these reactions one by one, extracted the involved substrates and cofactors and put them in the list of "starting compounds". We then applied all *the generalized reaction rules* simultaneously for three iterations to investigate can BNICE.ch replicate these reaction mechanisms in several steps. These identified steps represent a "multi-step reaction mechanism". The level of supervision in this approach was the same as described in Table 3.1.

The results of these two studies allowed us to explore *all possible enzymatic reactions (known and novel) between KEGG compounds*.

Table 3-1: The level of supervisions that is applied in this study for the generalized reaction rules, compounds and reactions.

Supervision level		Input constraint		
	Example	**Reaction rules**	**Compounds**	**Reactions**
Unsupervised	-		Unlimited	Unlimited
Database	KEGG	291*2	14'549	
Organism				
Network				
Pathway				

3.2.3 BridgIt Analysis

The concept behind the development of BridgIt framework is thoroughly discussed in Chapter 2. In this section, we applied the BridgIt method to assess the chemical similarity of the *hypothetical generated reactions* to the known KEGG reactions and to propose candidate genetic sequences for them.

Based on the assumption that enzymes with similar structure catalyze similar reactions, the BridgIt algorithm proposes a genetic sequence for the novel reactions, together with a score that indicates the similarity of the novel reaction to the known reaction. BridgIt also indicates in which organism the similar enzyme can be found, which is important in metabolic engineering studies.

We applied the four different modes of BridgIt as discussed in Chapter 2:

i. BridgIt_Cofactors
ii. BridgIt_EC
iii. BridgIt_CO_EC
iv. BridgIt_original

And we investigated the impact of the introduced features on the performance of BridgIt and its predictions.

3.3 Results and discussions

We generated a repository of all possible enzymatic reactions between KEGG compounds. This collection is a valuable source of information for the biochemical and biological studies and its characteristics and significance will be presented and discussed in the following sections.

3.3.1 KEGG reactions covered by BNICE

~ 90% of all KEGG reactions were reconstructed using the generalized reaction rules of BNICE.ch. There are 1'261 reactions in KEGG database that have no assigned EC number, which means that there is no known reaction mechanism for these reactions. Since BNICE reports the EC number up to the third level, we can propose EC classification and therefore reaction mechanisms for some of the reactions missing the EC number. We could identify the first three EC identifiers for 178 KEGG reactions. For 134 reactions the classification was unambiguous, meaning that there was only one suggested EC classification. For remaining 44 reactions two or more EC suggestions were reported. In most of the cases the suggestions were similar, i.e. out of the 46 ambiguous classifications only 10 reported different classifiers for the first EC level. Table 3.2 shows an example of KEGG reactions without corresponding EC number together with the third level EC number that BNICE.ch proposes for these reactions.

KEGG ID	Equation	Reaction rules	Suggested EC classification
R00091	C00003+(2)C00010<=>C00004+C00080+C02015	|1.8.1B1(rev)|1.8.1B1	1.8.1.-
R00270	C00014+C00026<=>C00001+C05572	|3.5.1A1(rev)|3.5.1A1|6.3.1A1(rev)|6.3.1A1	3.5.1.- | 6.3.1.-

Table 3-2: Two examples of KEGG reactions without a corresponding EC numbers and the suggested EC number by BNICE.ch. An unambiguous EC suggestion is in white, whereas reactions with two or more suggested EC numbers are highlighted in red.

The entire list of KEGG reactions with EC suggestions can be found in the Appendix, Table A1.

3.3.2 Multi-step reactions

An example of a multi-step reaction that is reconstructed with BNICE.ch is shown in Figure 3.1. Part A shows the description of a KEGG reaction with unidentified enzyme and mechanism. In Part B, using BNICE.ch, we propose a two-step reaction mechanism to carry out the biotransformation. The first step is a known KEGG reaction, but in the reverse direction and the second step is a novel reaction. Interestingly, for the first step we propose two reaction mechanism, one is the same as the known reaction on the reverse direction (3.5.1A1(rev)) and the other one is a class 6 enzyme that BNICE.ch proposes for this reaction (6.3.1.A1).

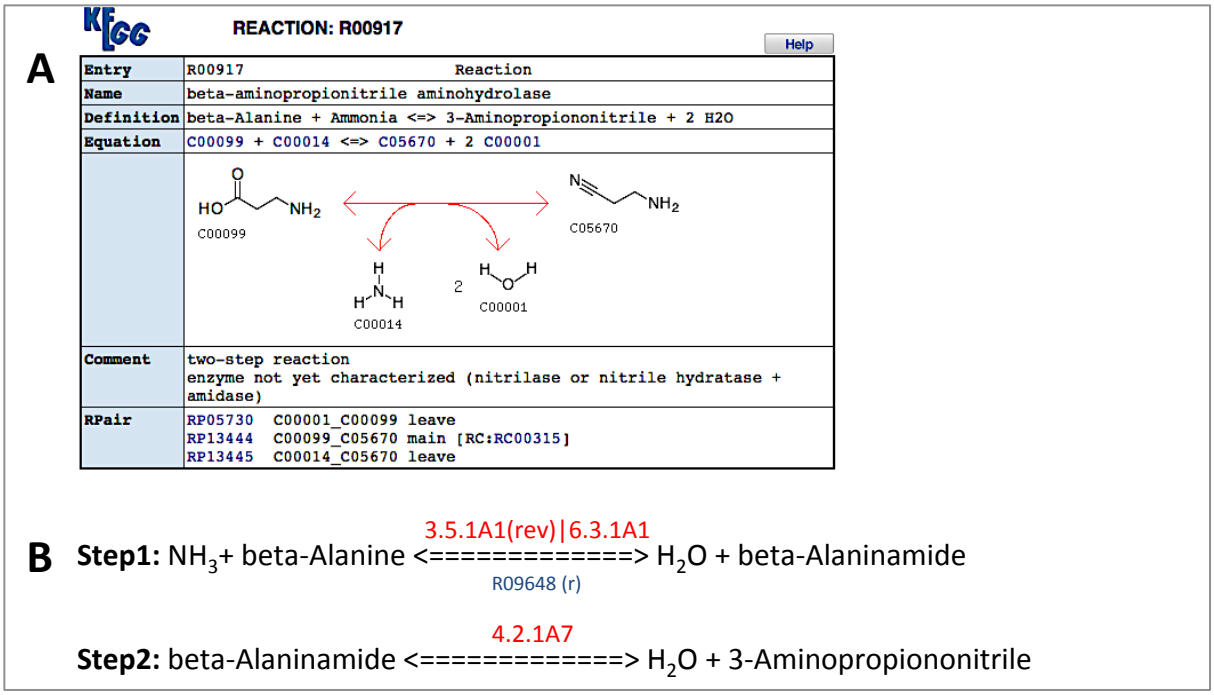


Figure 3-1: A two-step reaction of KEGG database that do not have a corresponding enzyme for catalyzing it (A) and using BNICE.ch, in (B) we proposed a two-steps reactions that has one known KEGG reactions (step 1) and one novel reaction (step 2) to describe the biotransformation.

3.3.3 Exploring the potential enzymatic capacity of biological compounds

Besides reconstructing ~90% of the KEGG reactions, BNICE.ch have also generated ~123,000 theoretical enzymatic reactions. This collection of novel reactions was generated by applying the whole set of 582 forward and reverse reaction rules against 14'549 KEGG compounds. This result indicated that there is a huge potential for

biological compounds to be interconverted by a relatively small number of reaction mechanisms.

3.3.4 Validating the action of generalized reaction rules

To validate the consistency of mechanisms of the generated hypothetical reactions with the known biochemistry, we investigated if any of the *novel* reactions that BNICE.ch generated on the basis of KEGG 2012 became *known* in the more recent versions of the KEGG database and we used KEGG 2014 for this comparison.

Between the two releases KEGG 2012 and KEGG 2014, 691 reactions were added to KEGG reaction database. From this set of added reactions we removed all incomplete or unbalanced reactions, as well as the reactions that contained compounds with undefined structure. We also removed reactions that could not possibly be reconstructed by BNICE because they involved new compounds that were not present in KEGG 2012, but were added later to KEGG 2014. After this preprocessing, the remaining data set of new entries of KEGG 2014 contained 236 reactions.

We then generated the set of novel reactions on the basis of KEGG 2012 reaction and compound databases and compared it with the 236 new entries of KEGG 2014. Interestingly, we predicted 81 of these 236 reactions. Even more strikingly, for 67 reactions the EC numbers that have been assigned by BNICE.ch matched the EC numbers in KEGG 2014. For 8 reactions BNICE.ch proposed an alternative classification, whereas for the remaining 6 reactions KEGG 2014 was lacking an EC identifier and BNICE.ch proposed an EC number up to the 3rd level for these reactions. The list of 81 reactions with their EC numbers is presented in Appendix Table A2.

3.3.5 BridgIT analysis of novel reactions

We performed 4 studies using 4 different methods of BridgIt to analyze the structural similarity of ~123,000 novel reactions predicted by BNICE.ch to the known KEGG reactions. We compared BridgIt performance for two criteria:

i. The distribution of the highest similarity score that each method predicts (Figure 3.2)

ii. The closest KEGG reaction that the 4 methods assign to the novel reactions. For this criterion, we compared combinations of pairs of the 4 methods across six different EC classes. (Figure 3.3)

The distribution of the highest score was quite different across 4 methods (Figure 3.2). As we applied 4 methods in the order of increasing constraints (i.e. from BridgIt_original to BridgIt_CO_EC) the number of high scores (close to 1) gradually decreased. The BridgIt_original method predicted high scores (close to 1) for the largest number of the novel reactions, whereas BridgIT_CO_EC predicted the smallest number of high scores. The accuracy of these methods can be tested against the BLAST similarity scores of these reactions. However, the preliminary results and verifications of the individual examples suggested that BridgIT_CO_EC had the higher accuracy.

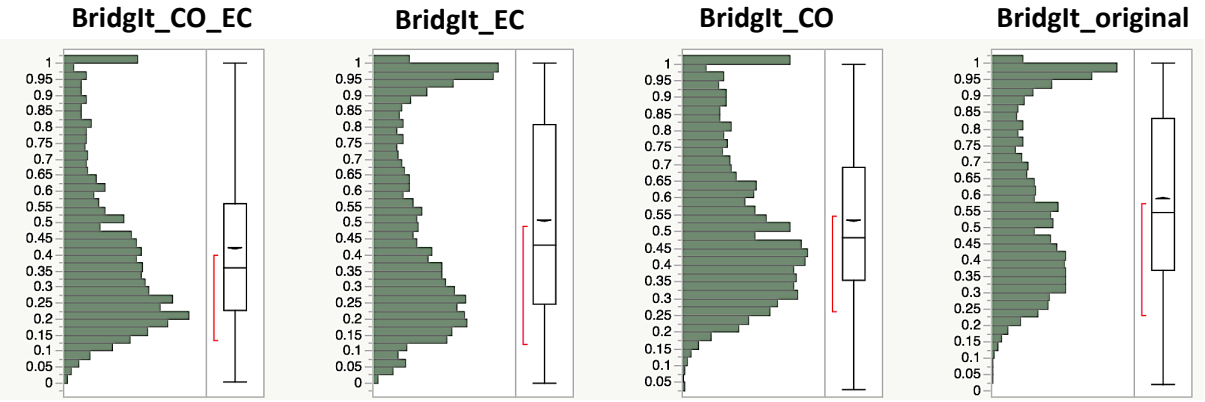


Figure 3-2: the distribution of the highest similarity score that different methods of BridgIt assign to the 123,000 novel reactions.

We compared combinations by pairs of 4 BridgeIt methods across six different EC classes to gain more insight about their performance (Figure 3.3). For example, the comparison of BridgIT_EC_CO and BridgIt_EC methods indicated that for the EC class 1, their predictions were ~30% similar, whereas for the EC class 4 their predictions coincided in more than 70% (Figure 3.3).

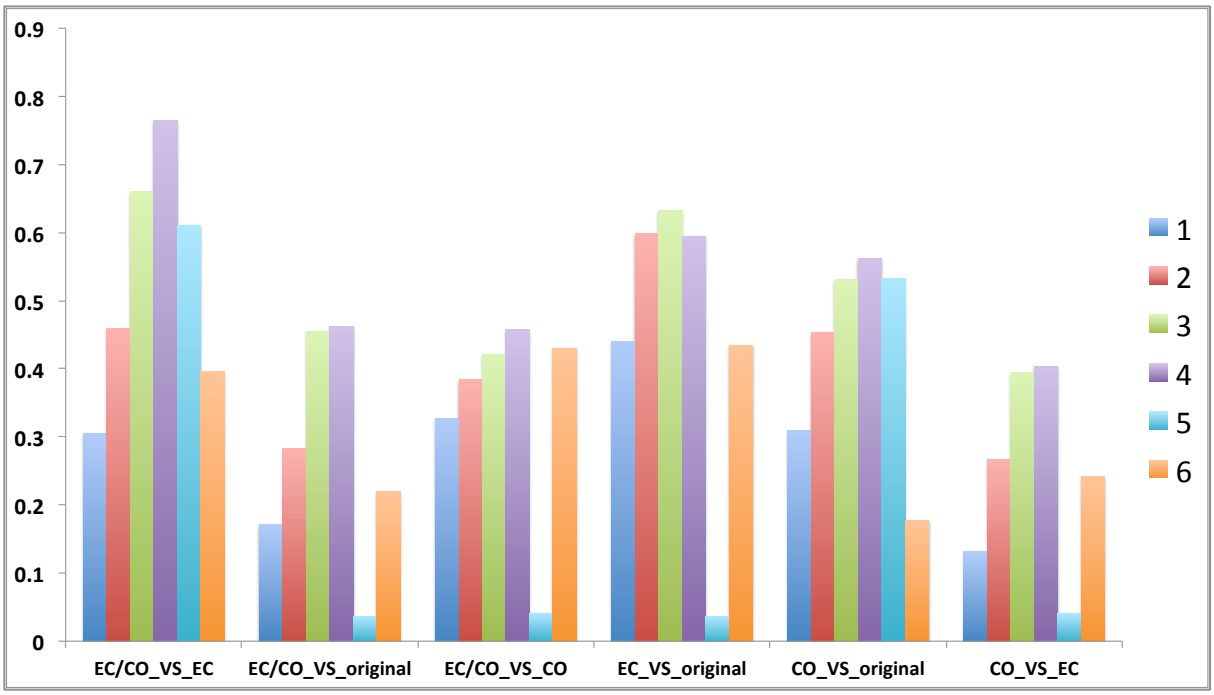


Figure 3-3: Comparing the similarity of the closet KEGG reaction that each method predicts for novel reactions. The first comparison is between BridgIT_EC_CO and BridgIt_EC methods (EC/CO_VS_EC). The color coding of the bars is according to the EC level 1 number.

3.4 Conclusions

We proposed a large collection of novel reactions along with their EC identifiers up to the 3rd level and candidate enzymes that may catalyze these novel reactions.

This high number of possible reactions connecting KEGG compounds is a very important finding, especially if we consider that BNICE.ch truly predicts biologically important reactions as witnessed by a correct prediction of 81 new reactions of the KEGG 2014 database. This study also illustrated that the current knowledge about metabolic reactions is far from being complete: to our best knowledge there is no database that accounts for such theoretically possible reactions, and we are first to propose a data collection of systematically generated novel reactions. “Atlas of Biochemistry” provides a valuable source of information for those who build and analyze metabolic models as well as for metabolic engineers searching for new biosynthesis or biodegradation pathways. “Atlas of Biochemistry” can be consulted at the following website: lcsb.epfl.ch/database

Chapter 4

NICELips

How to bridge the gap between the enoous available knowledge from lipidomics studies and the limited knowledge on lipid metabolism?

SUMMARY

In this chapter, we introduce a computational approach, called NICELips (Network Integrated Computational Explorer for Lipidomics), based on the formulation of generalized enzymatic reaction rules for lipid metabolism. Our approach employs the generalized rules to postulate novel pathways of lipid metabolism. It further integrates all discovered lipids in biological networks of enzymatic reactions that consist their biosynthesis and biodegradation pathways.

We illustrate the utility of our approach through a case study of bis(monoacylglycero)phosphate (BMP), a biologically important glycerophospholipid with immature synthesis and catabolic route(s). Using NICELips, we were able to propose various synthesis and degradation pathways for this compound and several other lipids with unknown metabolism like BMP, and in addition several alternative novel biosynthesis and biodegradation pathways for lipids with known metabolism. NICELips has potential applications in designing therapeutic interventions for lipid-associated disorders and in the metabolic engineering of model organisms for improving the bio-based production of lipid-derived fuels and chemicals.

4.1 Introduction

One of the most functionally and structurally important classes of lipids is glycerophospholipids. They constitute the majority of the common lipid class called phospholipids and can be found in almost all organisms as they are the building blocks of cell membranes [171,172]. Glycerophospholipids contain a polar head and a glycerol core with fatty acids attached to the glycerol moiety [173]. They are subdivided into distinct subclasses, based on the nature of the polar head group at the sn-3 position of the glycerol backbone [174] (Figure 4.1).

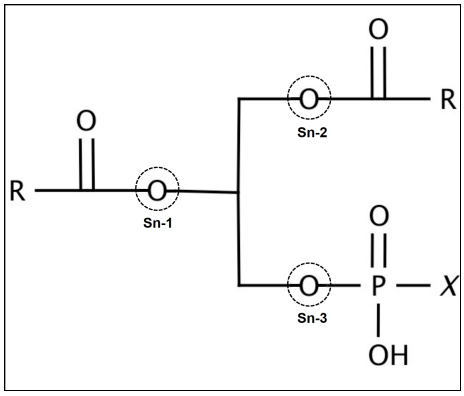


Figure 4-1: General structure of glycerophospholipids. In most glycerophospholipids the phosphate (on sn-3 position) is one of the following polar head groups: serine, choline, ethanolamine, glycerol, or inositol (designated X at right). The Acyl chains are shown with "R" in the n-2 and sn-1 positions and can have different length and degrees of saturation.

One of the major limitations in lipidomics and more generally in metabolomics experiments is the identification of unknown compounds due to the lack of comprehensive integrated reference databases [21,175,176].

Until recently, there were few specialized databases focusing on lipids analysis and classifications. In 2007, LIPID MAPS Structure Database (LMSD) became available [175]. LMSD is a comprehensive database for lipid structures that currently contains over 40,000 different classified structures and their corresponding physicochemical information. However, a comprehensive database of lipid structure is not enough to fully understand their multiple biological roles in cell biology and pathology. To further clarify their functions and the enzymes related to their metabolism, it is essential to organize them in the context of biological pathways and derive their associations and

interactions with enzymes and other lipids.

For the representation of lipids pathways, KEGG (Kyoto Encyclopedia of Genes and Genomes) [60] is the most comprehensive database available for small molecules and contains biological pathway maps for different parts of lipid metabolism. However, lipid biological pathways in KEGG are limited to general lipid species and do not include all the lipid structures available through LMSD. Therefore, the growth rate of these two databases is different and this results in the creation of a big knowledge gap between them.

The aforementioned needs and limitations in lipids bioinformatics motivated us to develop a computational framework, NICELips (Network Integrated Computational Explorer for Lipidomics), to generate associations between KEGG and LMSD databases and to enrich our knowledge of lipid metabolism. NICELips consists of several components integrated into a workflow and it is the first tool to provide an efficient and consistent procedure for linking lipid compound databases, such as LMSD, with pathway databases, such as KEGG. The central component involves the generation and reconstruction of lipid structures integrated in metabolic reactions. Within this component we identify all the known enzymatic reactions of lipid metabolism in KEGG database, and we formulate the generalized reaction rules for lipid reactions based on the molecular signatures of known enzymatic reactions. We then use BNICE.ch framework to apply the generalized reaction rules through *three different schemes* that differ on the level of supervision procedure (detailed description of the method is discussed in Chapter 2).

In this work we focused on the "Glycerophospholipid metabolism". In each of the three different schemes, we investigated a specific research question about the glycerophospholipid structures and metabolism.

In the *first scheme*, we reconstructed a comprehensive network of glycerophospholipid metabolism that includes all the known reported structures and reactions in their corresponding KEGG map. In addition we incorporated many *novel* compounds and reactions that have not been previously reported in databases. In order to facilitate the identification of novel compounds and assess the biochemical similarity between the novel and known compounds, we adapted and implemented a subgraph isomorphism

classification algorithm to classify these compounds based on the existing classification scheme for glycerophospholipids. The structural similarity among the members provides an insight on the functions of the novel compounds and enzyme specificity involved in their biosynthetic and catabolic pathways. Our classification method is also coupled with a scoring algorithm in order to assess the similarity of compounds in each group with the introduced general substructural pattern. The scoring algorithm assigns a value between 0 and 1 to each compound, and helps us to assign priority in the study of novel compounds. If we allow all the novel compounds to be kept in the generated reactions network, it limits the number of iterations that we can run BNICE.ch since it generates too many compounds and causes the combinatorial explosion.

In the *second scheme*, we repeated the same procedure as in the first scheme, but we constrained the "level of compound supervision" to only the KEGG compounds, in order to explore all the possible reactions between only the known lipid structures.

In the *third mode*, we applied the retrobiosynthesis algorithm of BNICE.ch that uses the reaction rules against LMSD structures to identify their metabolic and catabolic reaction pathways. We illustrate the efficacy and usefulness of our approach in the study of bis(monoacylglycero)phosphate (BMP) metabolism. BMP has two glycerol subunits linked by a phosphodiester group and it is a structural isomer of phosphatidylglycerol (PG). The endosomes are highly enriched in BMP, where it can amount up to 70% of the total phospholipids of the endosomal membrane [177]. BMP is important for the structural and functional integrity of the late endosomes. Interestingly, BMP is also a unique lipid due to its stereochemical configuration different from that of other animal glycerophospholipids. Despite numerous studies, we still miss essential knowledge concerning its properties, and biosynthetic and catabolic pathways. Based on the experimental evidence, BMP is synthesized from its structural isomer, (PG). After the removal of one fatty acid from the sn-2 position, lysophosphatidylglycerol is produced as next intermediate that then undergoes a transacylation reaction [178]. The results of the retrosynthesis experiment suggest various synthesis and degradation pathways for this compound.

Our studies here demonstrate how NICELips can provide a full overview of all lipid species in the cell, and particularly in the context of metabolic pathways that comprise

all the chemical interactions and transformations between lipid compounds and enzymes. The results of this work have important implications for discovering novel therapeutic approaches for lipid-associated disorders, through proposing novel biosynthetic and biodegradation pathways to alternate the metabolism of genetically defective lipids. On the other hand, exploring the entire space of feasible reactions in lipid metabolism will open up opportunities for generating *de novo* reactions to design and engineer new strains for the bio-based production of lipids-derived fuels and chemicals.

4.2 Methods

The development of the NICELips framework is an extension of BNICE.ch [72] tailored for lipid metabolism. NICELips specifically consists of two main algorithms of BNICE.ch and several auxiliary functions for further analysis of the obtained results. BNICE.ch methodologies are extensively discussed in Chapter 2.

The two main algorithms of BNICE.ch are applied in this study:

(i) *Forward network reconstruction algorithm*

To reconstruct the KEGG glycerophospholipid metabolism using specific generalized reaction rules relevant to this pathway

(ii) *Retrobiosynthesis algorithm*

To integrate the lipid structures from LMSD database into metabolic pathways and link them through known and hypothetical reactions by using the reaction rules.

A flowchart of the different steps of the NICELips framework is illustrated in Figure 4.2 and will be discussed in the following sections.

4.2.1 Generalized reaction rules in glycerophospholipid metabolism

The KEGG pathway of glycerophospholipid metabolism was used as a reference to develop the generalized enzymatic reaction rules for this study.

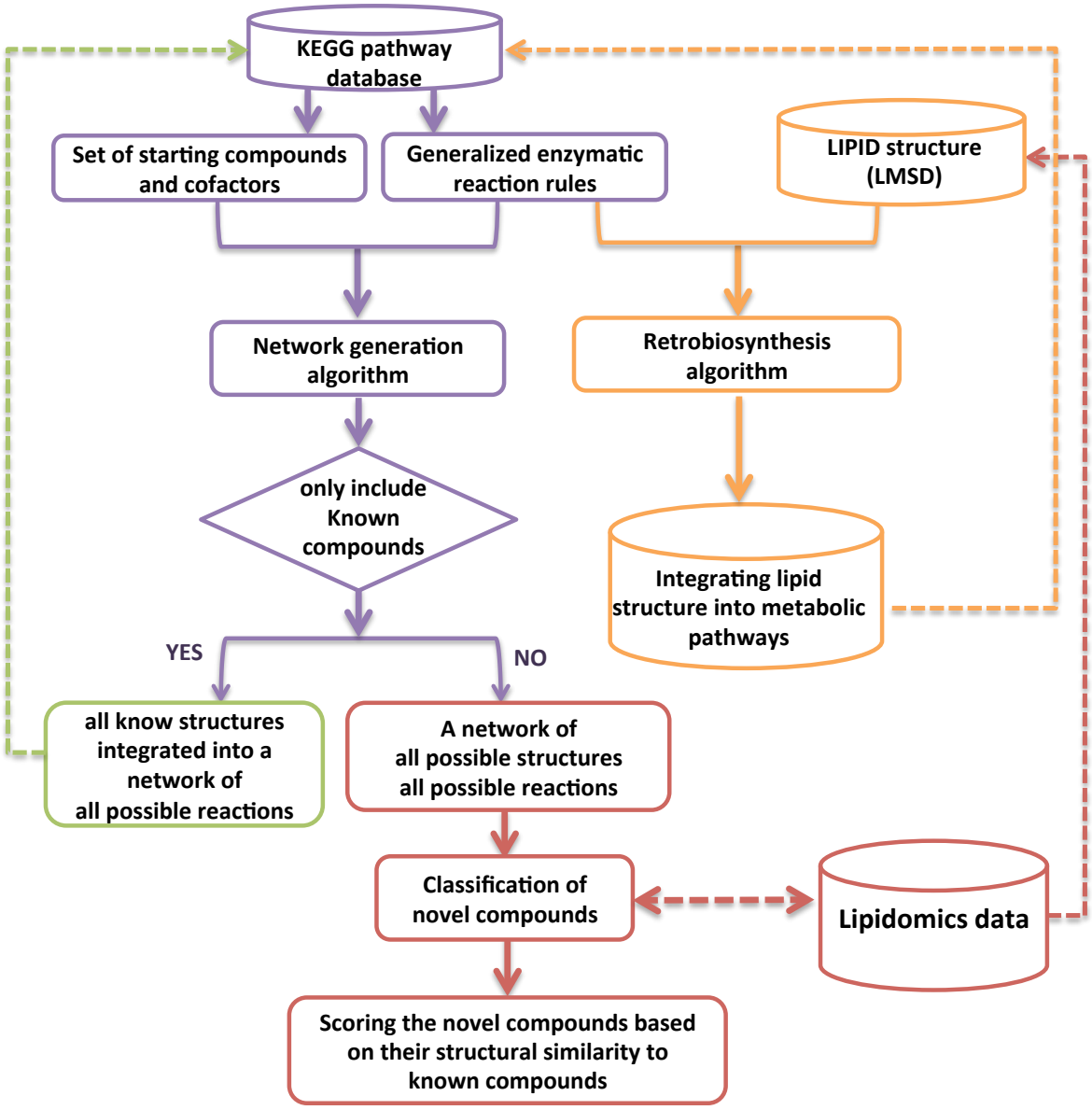


Figure 4-2: The NICELips framework for investigating lipid metabolism using three different schemes. All the three schemes start with a KEGG pathway that we want to study (glycerophospholipid metabolism in this study). We extract the generalized enzyme reaction rules relevant for the pathway under study and we reconstruct different levels of reactions networks based on the parameter we chose. Scheme 1 is highlighted in red, scheme 2 in green and scheme 3 in orange.

To clarify the process of developing the generalized reaction rules and their characteristics, we describe here an example of the process of formulating a reaction rule for the enzymes of the class EC 2.7.8 that are present in the KEGG pathway of glycerophospholipid metabolism.

Table 4-1: The set of enzyme rules involved in known glycerophospholipid pathways, in the parenthesis the specific reaction rules used in this study are shown. The alphanumeric classification of the rules corresponds to differences in cofactors and substrate structure (e.g., linear vs. cyclical structures). For each rule an example reaction with associated EC nomenclature and the required cofactors are given.

Enzyme rule	Reaction name	EC	Required cofactor
1.1.1 (1.1.1A1)	Oxidoreductases with NAD^+ or $NADP^+$ as acceptor	1.1.1.8	H^+, NAD^+ or $NADP^+$
2.1.1 (2.1.1A1)	Methyltransferases	2.1.1.103	S-Adenosyl-L-methionine
2.3.1 (2.3.1F1)	Transferring groups other than aminoacyl groups	2.3.1.15	Acyl-CoA
2.7.1 (2.7.1A1)	Phosphotransferases with an alcohol group as acceptor	2.7.1.32	ATP
2.7.7 (2.7.7A1)	Nucleotidyltransferases	2.7.7.41	Cytidine triphosphate
2.7.8 (2.7.8A1)	Transferases for other substituted phosphate groups	2.7.8.5	__
3.1.1 (3.1.1A1)	Carboxylic-ester hydrolases	3.1.1.52	Water, O_2
3.1.3 (3.1.3A2)	Phosphoric-monoester hydrolases	3.1.3.27	Phosphate
3.1.4 (3.1.4A1)	Phosphoric-diester hydrolases	3.1.4.3	Water
3.6.1 (3.6.1A1)	In phosphorus-containing anhydrides	3.6.1.26	Water
4.1.1 (4.1.1A3)	Carboxy-lyases	4.1.1.65	CO_2
4.3.1 (4.3.1A3)	Ammonia-lyases	4.3.1.7	Ammonia

In developing the rule 2.7.8, we studied all the 4th level reactions in EC 2.7.8 class (designated as: "Transferases for other substituted phosphate groups") in which there are 37, 4th level reactions (numbered from 2.7.8.1 to 2.7.8.37). The 37 4th level EC numbers correspond to 51 specific KEGG reaction IDs. By carefully investigating their mechanism, we found out that in 24 out of 51 reactions, the reactive sites of the substrates share common structures. Consequently, we developed the general reaction rule 2.7.8A1 for replicating the 24 reactions having exact match of reactive sites of their substrates.

Out of these 24 KEGG reactions, 11 reactions belong to the glycerophospholipid metabolism. For the remaining 27 reactions we further developed 2 additional rules (2.7.8A2 and 2.7.8B1), which do not appear in the glycerophospholid metabolism. The joined set of the 2.7.8 rules is able to reproduce all the currently known reactions that could be classified as members of the EC 2.7.8 class. Many rules require specific co-factors or other small molecules for their function (Table 4.1).

In order to reproduce all the reactions involved in the KEGG pathway of glycerophospholipid metabolism, we found that at least 12 generalized enzyme reaction rules were required (Table 4.1).

4.2.2 Forward network generation within NICELips

We employ the forward network generation algorithm to replicate the known pathway of glycerophospholipid metabolism as found in KEGG. Next, we apply an automatic search into KEGG database to identify which of the generated compounds and reactions are novel and which already exist in KEGG.

The initial set of starting compounds consisted of 4 core compounds (sn-glycerol-3-phosphate, serine, choline, and inositol) and 12 cofactors (NADH, NAD^+, H^+, Bicarbonate, CO_2, O_2, Water, Cytidine triphosphate (CTP), Phosphate, Acyl-CoA, S-Adenosylmethionine and ammonia), and the set of the 12 generalized reaction rules (Table 4.1).

We ran the algorithm with two different levels of supervisions. In the **first approach** we investigated the glycerophospholipid metabolism with supervision *only* on the generalized enzyme rules that describe the known glycerophospholipid pathway in KEGG. As for the search space of compounds and reactions, we didn't apply any constraint (Table 4.2). We wanted to understand after how many number of iterations the combinatorial explosion occurs for the compounds (known and novel) and for the reactions (known and novel).

Table 4-2: Applied levels of supervision for the analysis of the Scheme 1.

Supervision level		Input constraint		
	Example	Reaction rules	Compounds	Reactions
Unsupervised	-		Unlimited	Unlimited
Database				
Organism				
Network				
Pathway	Glycerophosphplipid	12*2	33	69

Since one of the main objectives of NICELips is to make associations between LMSD compounds and known metabolites found in KEGG database, we do not include novel compounds in the **second approach.** We investigated how the reaction network of glycerophospholipid will evolve if we explored all the possible reactions between known compounds of KEGG. Will it converge after a certain point?

To answer this question, we applied, the supervision on the level of compounds rather than the supervision of the generalized reaction rules. As for the search space of the compounds, we only allowed the KEGG compounds to be included in our network (Table 4.3).

Table 4-3: Applied levels of supervision for the analysis of the Scheme 2&3.

Supervision level		Input constraint		
	Example	Reaction rules	Compounds	Reactions
Unsupervised	-		Unlimited	Unlimited
Database	KEGG		14'549	5914
Organism				
Network				
Pathway	Glycerophosphplipid	12*2	33	69

4.2.2.1 Automatic classification and scoring

For the rapid and automatic classification of the compounds dataset generated by NICELips, we established a subgraph isomorphism algorithm, coupled with a scoring algorithm. Exploiting the fact that NICELips uses a graph representation of the compound structures, we used a chemistry-based approach by introducing a

classification algorithm based on subgraph isomorphism, in which the novel compounds are classified through comparison with characteristic substructures of these subclasses. More precisely, we used the existing classification scheme for glycerophospholipids in which seven main classes are distinguished by a base structure in each class. Identifying the key features of the base structures, we generated template (sub)graphs (substructures) for different classes and used them to classify thousands of novel compounds predicted with NICELips, using our proposed algorithm.

Several methods for (sub)graph matching are found in literature [179-181]. We adapted and used the efficient VF2 (sub)graph matching algorithm introduced in [182]. In (sub)graph isomorphism literature, the VF2 algorithm [183] is known to be one of the fastest and most accurate algorithms proposed so far. Its effectiveness is shown especially when applied to large graphs, and it is well established that this algorithm outperforms other existing subgraph isomorphism algorithms such as Ullmann, SD, Nauty and VF algorithms. Basically, a subgraph-matching algorithm is provided with a subgraph G_y of base structure y (i.e., any of the substructures depicted in Table 3), and a graph G_X of the compound X found using NICELips. The VF2 algorithm searches the structure of G_X, and if $G_Y \subset G_X$ (i.e., if graph G_X contains the subgraph G_y), places the compound X in the class defined by the substructure Y.

Using this method, based on the structural similarity, all novel compounds can be classified into the glycerophospholipid subclasses.

The analysis of the large number of novel compounds was facilitated by the application of a scoring algorithm. In this algorithm, for each class we introduced the key feature of the class (characteristic substructures) as the reference structure and the other members of the class were ranked according to their similarity to the reference structure of the class. The scoring is based on 2D similarity measures. Each compound is coded in a vector, called fingerprint, which codes the characteristic of the structure of a compound. Then, based on the comparison between two fingerprints, a similarity coefficient is computed, which gives us a similarity score. Different methods exist for creating fingerprints and calculating the similarity coefficient. We hereby elaborate on our choice of the two methods. For the similarity coefficient, the choice of the Tanimoto coefficient (also called Jaccard statistic) is a recommended choice in chemical

information by various studies. Therefore, we used this measure of similarity in our work.

The choice of the fingerprint method has to be deduced depending on its suitability to represent structural information for a set of compounds of interest. The Wild and Blankey [184] study evaluated the suitability of different 2D fingerprints to represent structural information. They conclude that fingerprints that use a predefined dictionary of structural features outperform fingerprints derived from structural paths in compounds when comparison is done to larger data sets with less well-defined clusters. However, fingerprints that encode structures obtain better results to discriminate the differences between closely related series. After performing tests on different molecules insight classes and across classes, we use for the scoring MACCS fingerprint. These well-established structural keys are encoded using a set of 166 public keys defined in the MDL documentation, implemented in OpenBabel [185]. The MACCS method is a valuable tool in the selection and prioritization of compounds from large compound collections.

4.2.3 Retrobiosynthesis algorithm within NICELips

In the **third** approach, we applied the retrobiosynthesis algorithm for LMSD glycerophospholipids. We validated the efficiency of our approach for linking KEGG and LMSD databases through the study of a biologically important lipid compound, bis(monoacylglycero)phosphate (BMP), which does not exist in KEGG database but exists in LMSD. The level of supervision applied in this study is exactly as Table 4.3 shows.

4.2.4 Pathway enumeration algorithm

To take full advantage of the wealth of information generated in both approaches, we apply our pathway reconstruction algorithm to analyze and evaluate the generated compounds and pathways. The pathway search algorithm identifies all linear pathways between a defined starting compound and target compound, and provides insight for the pathways required for the synthesis and catabolism of different lipid compounds.

4.2.5 Thermodynamic studies of the generated pathway

In order to prune the set of novel synthesis and catabolic pathways and to evaluate their relative feasibility, we performed a thermodynamic feasibility analysis based on the thermodynamics-based metabolic flux analysis (TMFA/TFBA) [47,186,187]. The standard Gibbs free energy of reactions is estimated using the group contribution method for biological compounds [48,188].

4.3 Results and discussion

We used NICELips to study the important class of lipids associated with the glycerophospholipid metabolism. One of the main challenges is the construction of a full comprehensive database of compounds, reactions and pathways for lipid metabolism. The huge gap between the two most commonly used databases for lipids, KEGG and LMSD, is growing rapidly due to advancement of mass spectrometry methods. The ability of our approach to generate and assemble a complete virtual database of lipid compounds and pathways, allows the concurrent filling of the gap between the two databases.

We applied NICELips framework to:

(i) explore the diversity of structures in glycerophospholipids and classify the novel generated compounds, (ii) generate a reaction network for glycerophospholipid metabolism to Expand KEGG metabolic pathway and (iii) link KEGG and LMSD databases through the study of a biologically important lipid compound, bis(monoacylglycero)phosphate (BMP), which doesn't exist in KEGG database but exists in LMSD. In the following paragraphs we explore the insight that we gained from NICELips in each of these applications.

4.3.1 Exploring glycerophospholipid metabolism

We explored the diversity of structures and reactions in glycerophospholipid metabolism by applying NICELips without any constraints on the search space for compounds and reactions. This resulted in the generation of a network of all possible compounds and reactions based on the enzymatic rules of glycerophospholipids. After nine iterations the produced network includes 4,497 compounds and 20,874 reactions

4.3.1.1 NICELips vs. KEGG and LMSD

We compared the generated results (compounds and reactions) with KEGG database in which there are nomenclatures for both compounds in the form of their chemical formulas and structures and reactions. The KEGG glycerophospholipid pathway includes 33 compounds and 69 reactions. In our generated dataset, there are 74 compounds and 68 reactions which fully cover the 33 compounds and 69 reactions of the KEGG glycerophospholipid pathway (Table 4.4). There exist also other KEGG compounds and reactions not assigned in the glycerophospholipid map (specifically 25 compounds and 13 reactions), some of which are assigned to other metabolic maps, while few of them are not assigned to any metabolic map of the KEGG database. This result highlights an important utility of our framework, which is the postulation of enzyme-based hypothetical reactions that can link compounds to reactions and integrate them into existing metabolic maps.

In LMSD database, several fatty acid compositions have been described for lipids. The huge diversity in the LMSD database comes mainly from the variation of attached acyl chains "fatty acids" in terms of length, number and position of double bonds.

In our framework and in consistency with KEGG database, fatty acid substructures are all represented with a common group name (called "R"); thus we assume a unified substructure and do not consider the different modifications of the acyl chains in lipid structures. If one ignores the acyl chain variations, then LMSD contains a limited number of distinguishable structures of glycerophospholipids. More precisely, we could identify 62 unique structures in this database. We then compared our generated dataset with LMSD and we observed that our dataset includes 51 structures (82% out of the total unique structures in LMSD). The remaining structures that NICELips were not able to generate were due to the following reasons:

(i) The molecule contained head groups that required other starting compounds, such as sugars, not in the KEGG map, and

(ii) The reactions involved biotransformations that were not captured by our current set of generalized reaction rules.

Table 4-4: Statistics of compounds and reactions generated in the first scheme, after 9 iterations of NICELips algorithm.

Iteration	Reactions predicted (cumulative)	Of which in KEGG	Compounds predicted (cumulative)	Of which in KEGG
1	3	3	21	20
2	15	8	36	30
3	62	18	62	41
4	141	26	88	44
5	247	32	123	50
6	446	42	186	58
7	949	49	329	65
8	3,054	57	889	70
9	20874	68	4497	74
Total	20,874	68	4,497	74

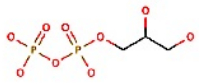
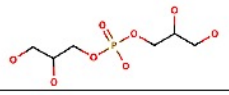
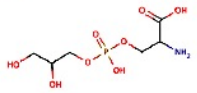
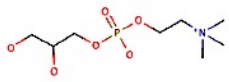
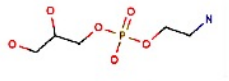
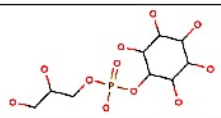
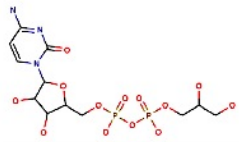
Given that the rules we used in this study are based on the enzyme reactions in known glycerophospholipid pathways, this finding suggests a broader class of enzymes are involved in glycerophospholipid metabolism. This is an observation that has not been acknowledged before in the classification of glycerophospholipid metabolism. Overall, these results demonstrate the very good match of our dataset with the data in LMSD.

4.3.1.2 Classification and scoring of the compound dataset

Unlike other classes of biomolecules that can be considered as mere permutations on a common and finite set of monomers (e.g., proteins and oligonucleotides), complex lipids, such as glycerophospholipids, are composites of a wide range of building blocks that can give rise to a huge array of combination. These combinations cannot be predicted by simple inspection of the precursor molecules because they are the product of a large number of biochemical transformations. The diversity in molecular structure was extremely higher in our datasets, since we explored computationally the diversity of glycerophospholipid structures using all the possible enzyme-reaction rules involved in the synthesis of these groups. Common glycerophospholipids of mammalian tissues can be classified according to the following head group components: 1) phosphate in Glyceropyrophosphates (PPA); 2) glycerol unit in Glycerophosphoglycerols (PG); 3) serine group in Glycerophosphoserines (PS); 4) choline in Glycerophosphocholines (PC); 5) ethanolamine in Glycerophosphoethanolamines (PE); 6) six-carbon sugar

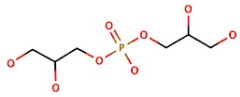
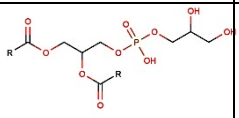
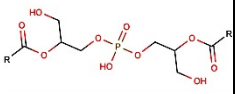
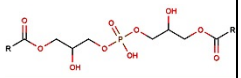
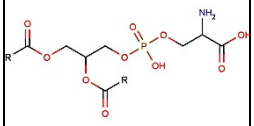
inositol in Glycerophosphoinositol (PI); 7) Cytidine 5'-diphosphate in CDP-Glycerols (CDP-Glycerol).

Table 4-5: Results of the classification of the novel compounds in 7 classes based on their structural similarities to the known lipid structures.

Class	Subgraph structure	# compounds generated (in KEGG, in LMSD)
PPA		302 (5,5)
PG		402 (6,11)
PS		124 (3,5)
PC		280 (4,4)
PE		692 (4,13)
PI		845 (4,6)
CDP-Glycerol		1086 (1,2)
Others		766 (6,5)

In order to tackle the diversity of the generated molecular structures, we first constructed a classifier to classify 4,497 compounds belonging to seven different structural classes, based on the chemical structures of the known compounds using the subgraph isomorphism technique described in section 5.2.2.1.

Table 4-6: A closer look the compounds of the PG class shows that two different stereoisomers of bis(monoacylglycero)phosphate is generated as a novel compound in this study and is correctly classified in the PG class, with a high similarity score of 0.84 .

Introduced pattern for scoring in PG class		
		
Results		
Name of compound	Structure	Similarity score
Phosphatidylglycerol		0.90
BMP 2,2		0.84
BMP 3,3'		0.84
Phosphatidylserine		0.69

We next compared the compounds in the generated datasets with those in KEGG and LMSD. In order to organize the groups in a controlled manner that will facilitate the comparison, we merged 13 classes of LMSD with the same head groups and came up with 7 different structural-based classes, plus one containing “other structures”, similar to LMSD. The inspection of the summary of classification results (Table 4.5) suggests that there are many structures in different classes not existing in KEGG, which either exist in other databases such as LMSD, or do not exist in any current lipid database.

In our method of classification we have devised for each compound a scoring value between 0 and 1. This scoring value shows the structural similarity between all the members of the group and the general pattern used for classification, while helping us to assign priority in the study of novel compounds. In other words, for each novel

compound, a score closer to 1 accounts for higher similarity to known compounds and higher likelihood for its existence. One can arbitrarily define threshold values for the score and select all compounds from a class, with a score beyond a particular threshold, for more analysis. Selecting compounds according to the scores provides us with a manageable number of novel compounds for further studies on biosynthesis and catabolic pathway association with these novel compounds. Hereby we show the results of our case study for a lipid compound that does not exist in KEGG: bis(monoacylglycero)phosphate (BMP) found in the PG class with the similarity score of 0.848. Examples of our scoring results are shown in Table 4.6.

4.3.2 Expansion of the glycerophospholipid KEGG pathway

The KEGG pathway map of glycerophospholipid includes 33 compounds and 69 reactions. After 20 iterations of our algorithm, the number of generated compounds converged to 383 (all KEGG compounds) and the number of generated reactions converged to 3787. Out of the 3787 generated reactions, 251 reactions corresponded to KEGG reactions as shown in Figure 4.3. Within the generated results, we recovered all the 33 compounds and 69 reactions as reported in the KEGG glycerophospholipids pathways. We believe that the fact that NICELips finds all the previously curated compounds and reactions in KEGG is an implicit validation of the NICELips framework for the study of lipid metabolism.

Interestingly, many of these KEGG compounds and reactions have not been previously associated with the glycerophospholipids pathway in KEGG database. This observation demonstrates how our method can now link the KEGG pathway of glycerophospholipids with the other KEGG pathways and also propose metabolic pathways for reactions that are not associated with any metabolic pathways in KEGG database.

Our results also proposed several shorter alternative pathways including novel steps for the synthesis of biologically important lipids in glycerophospholipid metabolism.

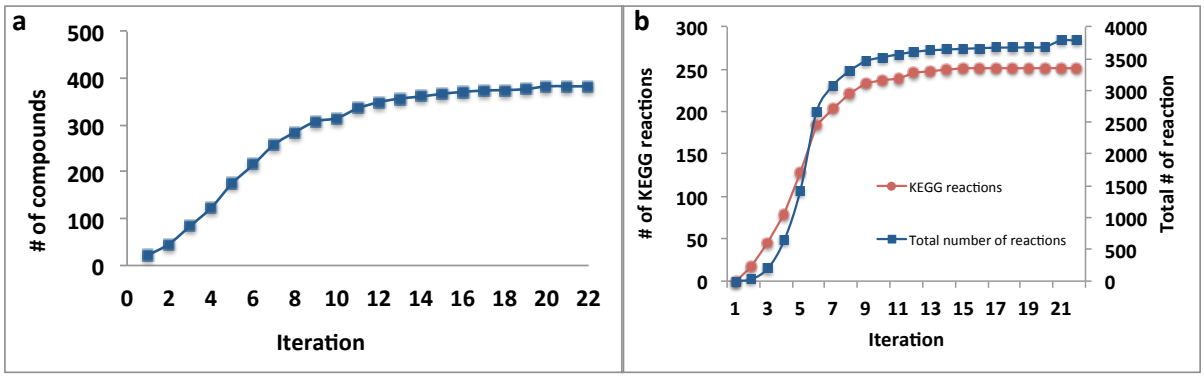


Figure 4-3: (a) Compounds generated through the reconstruction of glycerophospholipid pathway using NICELips. The number of compounds increases in each iteration. (b) Total number of reactions and KEGG reactions generated in each iteration.

4.3.2.1 Reconstruction and enumeration of novel pathways

We applied the pathway search algorithm to find alternative pathways for the synthesis and biodegradation of known lipids. We started with a defined substrate and a target lipid, and we applied the algorithm to reconstruct all linear connections between them. The starting compound we used in this study is sn-glycerol-3-phosphate, which is the substrate we also used for the reconstruction of the KEGG map of glycerophospholipids (see above). We chose as target compound the phosphatidylglycerol (PG), which is known to be the substrate of bis(monoacylglycero)phosphate (BMP) biosynthesis. There is only one known *de novo* pathway for PG biosynthesis which is a 4-step pathway starting from sn-glycerol-3-phosphate (Figure 4.4). Interestingly, we find hundreds of novel pathways, which are even shorter than the native pathway in KEGG database (representative shorter pathways are shown in Figure 4.4).

4.3.3 Retrobiosynthesis of bis(monoacylglycero)phosphate (BMP)

We aimed to find hypothetical metabolic pathways for lipids with unknown metabolism, such as BMP. Starting with a set of LMSD lipid compounds and our reaction rules curated based on KEGG reactions we applied the retrobiosynthesis algorithm in order to identify the pathways that link lipids in LMSD back to metabolic intermediates. We hereby show our results of the retrosynthesis approach for BMP.

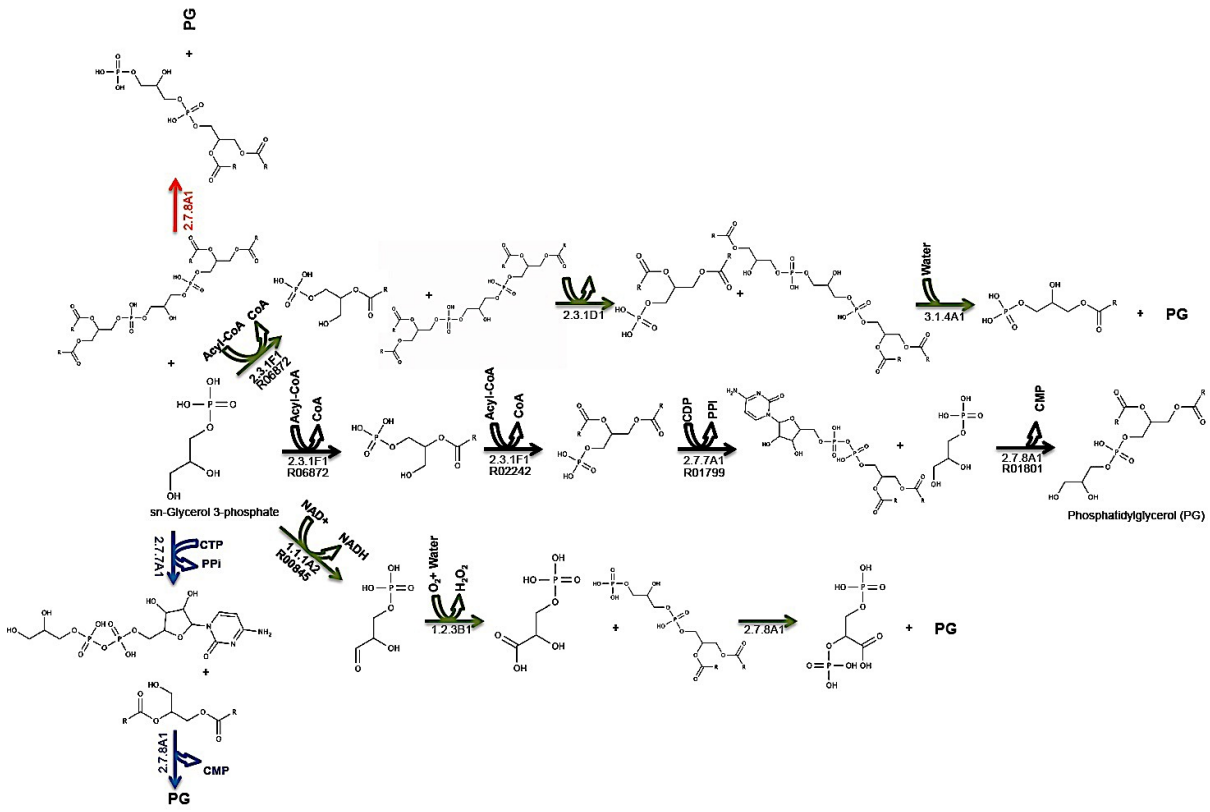


Figure 4-4: The native de novo pathway for PG biosynthesis as is found in KEGG is shown in black arrows. Based on our results we propose several alternative shorter pathways with length 1 (in red arrows), length 2 (in blue arrows) and length 3 (in green arrows). The pathways presented here are chosen based on the thermodynamics feasibility results and we present the most feasible pathways for each length.

BMP, the structural isomer of phosphatidylglycerol (PG) is one of the most intriguing structures of phospholipids discovered to date [189]. It exhibits an unusual sn-1:sn-1' stereo configuration, based on the position of the phosphate moiety in its two glycerol units. Based on the positions of fatty acids on the glycerol molecule, two different isomers have been reported for BMP, the 2,2'-dioleoyl form and the 3,3'-isomer.

We carried out 24 iterations of the retrosynthesis algorithm using the set of 12 reaction rules, starting compounds and cofactors (mentioned in section 2.2), plus BMP. After 24 iterations the number of compounds (all KEGG compounds) converged to 208 and the number of reactions converged to 505 (out of which 101 reactions are KEGG reactions) as shown in Figure 4.5. We found that the minimum number of steps for the synthesis of

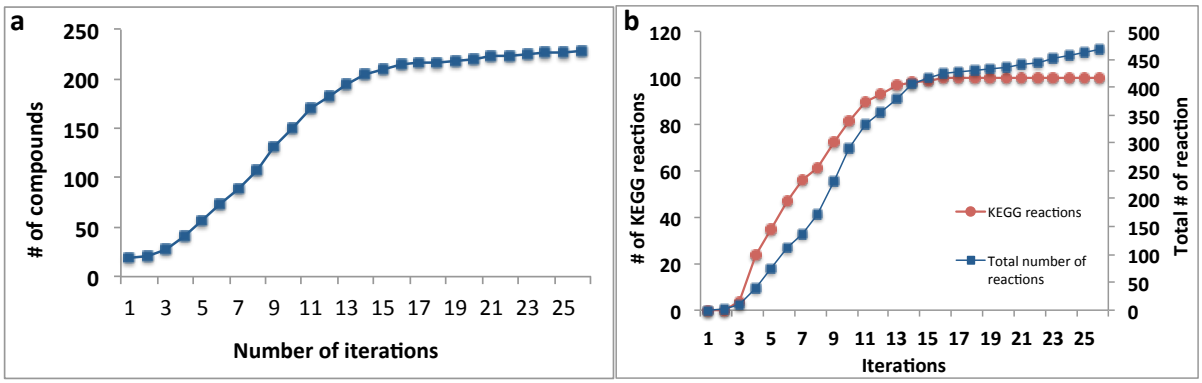


Figure 4-5: The results of the retrosynthesis algorithm applied for BMP. (a) shows the iterative generation of compounds and (b) shows the total number and KEGG reactions generated for each iteration.

BMP from PG is 4, and it involves 7 lipid compounds (Figure 4.6). Moreover, we found shorter pathways from other known lipid compounds that could lead to BMP. As shown in Figure 4.6, BMP is one step away from the known compounds reported in KEGG database. After a one-step reaction with water as a cofactor (a novel reaction that does not exist in KEGG), BMP is degraded to "2-Acyl-sn-glycerol 3-phosphate" and "2-Acylglycerol". "2-Acyl-sn-glycerol 3-phosphate" is a lipid compound associated with glycerophospholipid pathway in KEGG database. "2-Acylglycerol" is a KEGG compound which is not associated with any reported metabolic pathway.

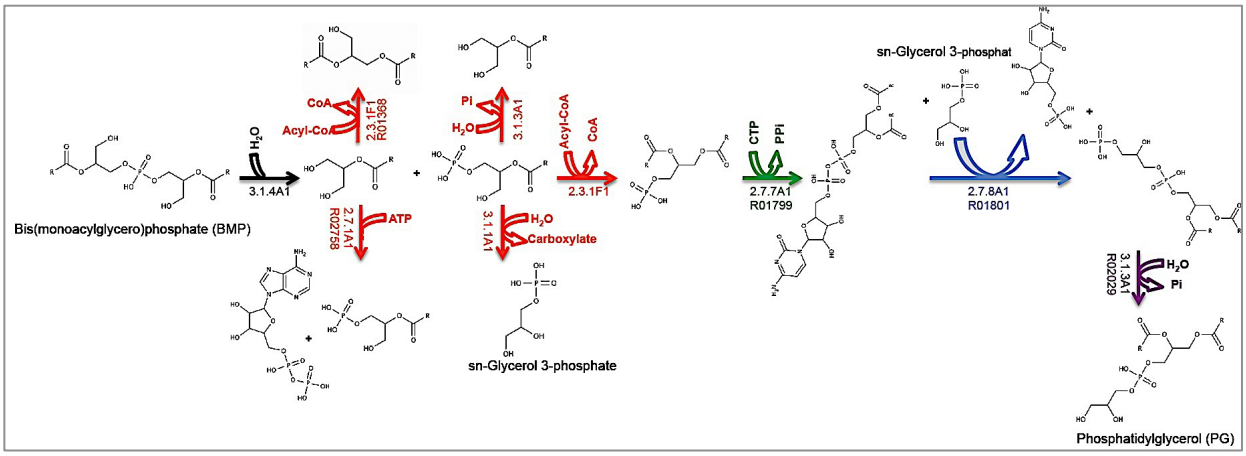


Figure 4-6: Out of hundreds of generated reactions for BMP metabolic network, hereby we present couples of thermodynamically feasible pathways. In the set of required generalized reaction rules for reconstruction of the KEGG pathway of glycerophospholipids, "3.1.4A1" is the only enzyme rule that can react with BMP (first iteration showed in black arrow in the figure). In the second iteration (showed in red arrows), sn-Glycerol 3-phosphate which is the substrate for PG biosynthesis in

glycerophospholipid metabolism is generated. In the third iteration (showed in green arrow), sn-Glycerol 3-phosphate is generated through another reaction. And finally, in the fifth iteration (showed in purple arrow), phosphatidylglycerol (PG), which is supposed to be the substrate for BMP biosynthesis is produced.

Interestingly our results propose a possible integration of this lipid to glycerophospholipid pathway. Thus, based on the results obtained, we are able to propose alternate candidate synthesis and degradation pathways for BMP with different lengths and biophysical properties.

We observed that not all of the LMSD compounds are one step away from KEGG compounds. As mentioned in section 4.2, we can exclude novel compounds from the generated network, and this results in smaller network that includes only the KEGG compounds. For the LMSD compounds that are not one step away from the KEGG compounds, we allow all novel and known compounds in each generation, and we obtain networks that have both novel and KEGG compounds. The novel compounds serve as intermediates to connect the LMSD compounds to the KEGG metabolic pathways.

4.3.4 Thermodynamic feasibility studies

We employed the group contribution method to estimate the standard Gibbs free energy ($\Delta°G_r$) for all the reactions generated using the NICELips framework (Table 2). $\Delta°G_r$ is estimated in the range of +28.49 KJ to -35.64 KJ for 93% of 3497 reactions in the reconstructed networks of glycerophospholipid metabolism. The $\Delta°G_r$ for 64% of 505 reactions in the networks of the retrosynthesis of BMP is estimated in the range of +13.37 KJ to -35.64 KJ. We were not able to estimate the energies for some of the reactions due to groups in the substrates or products for which thermodynamic information was not available.

To further analyze the predicted alternative pathways, we performed thermodynamics analysis for the results of the pathway search algorithm. $\Delta°G_r$ of the overall pathway can be computed by the sum of $\Delta°G_r$ of the individual reactions in the pathway. If $\Delta°G_r$ of the overall pathway is negative we consider it as a thermodynamically feasible pathway. According to this calculation, all the representative pathways in this study (Figure 4.4 &

Figure 4.6) were found to be thermodynamically feasible.

Table 4-7: A summary of results of thermoynamics feasibility studies for reconstructed network of glycerophospholipids and the analysis of the reaction network of retrosynthesis of BMP.

Reconstructed network of glycerophospholipid			**Reaction network of retrosynthesis of BMP**		
Range of ΔG_r	# of reactions	% in KEGG	Range of ΔG_r	# of reactions	% in KEGG
$\Delta G_r < 0$	664	44.6%	$\Delta G_r < 0$	154	43%
$\Delta G_r = 0$	2231	5.4%	$\Delta G_r = 0$	36	2%
$\Delta G_r > 0$	601	45.3%	$\Delta G_r > 0$	129	53%
ΔG_r is not estimated	291	4.7%	ΔG_r is not estimated	186	2%

We also observed that the standard Gibbs free energy for almost 50% of the KEGG reactions is positive (Table 4.7). This indicates that the concentrations of the substrates and products in the system play a critical role in determining the feasibility *in vivo* as the Gibbs free energy of reaction has to be negative in order for the reactions to operate in the direction of the synthesis of important compounds. However, this entails knowledge of *in vivo* concentration ranges and network thermodynamic feasibility analysis. Therefore, having positive values for the standard Gibbs energy of reactions of the network is not enough for discarding them as thermodynamically infeasible reactions and there are other parameters that should be taken into account.

4.4 Conclusions

Due to the importance of lipids in cellular physiology and pathology as well as to the recent interest of the biotechnology community for understanding their role as resources for alternative fuels and chemicals, the study of lipids has emerged as a major research area. A computational framework, NICELips, was developed to investigate lipid metabolism. In this work, we demonstrated the utility of NICELips for studying glycerophospholipid metabolism. Our results emphasize the wealth of novel lipid structures and biosynthesis and the abundance of catabolic reactions that are yet to be discovered. We demonstrated our approach by reporting an interesting lipid compound that does not appear in KEGG: the bis(monoacylglycero)phosphate (BMP).

By reconstructing the KEGG pathways of lipid metabolism, our results indicated several shorter alternative pathways for lipids synthesis than those found in KEGG. The significant advantage of NICELips is that all the generated compounds are automatically organized into pathways of reactions, i.e., we can easily identify the reactions in which the compounds have participated. This is a unique property that enables us to search for pathways based on the desired compounds, as a result of the generation of a network of all compounds and reactions.

The retrosynthesis analysis using NICELips links all the reported structures for lipids in LMSD to the metabolic pathways in KEGG, and thus offers a great potential for proposing synthesis and catabolic pathways for many lipids in LMSD.

Due to the large number of novel pathways discovered for different compounds, we should next develop criteria for identifying and ranking of pathways that are more feasible *in vivo*. Such ranking will further guide the research in targeted lipidomics for the experimental identification of pathways around lipids of interest, for which information about their synthesis and catabolic routes is still missing.

Chapter 5

SUPER *E. coli*

How to discover and assess the full enzymatic and metabolic capabilities of an organism?

SUMMARY

In this chapter, we used BNICE.ch for the first time in an organism specific study to search for novel reactions that could potentially exist in the metabolic network of *E. coli*. Using a given set of *E. coli* core metabolites and a set of known biotransformation rules that exist in *E. Coli* as inputs for BNICE.ch we generated Super *E. coli,* which is an extended metabolic network with full metabolic capabilities that *E. coli* can potentially provide. Super *E. coli* captures all the known *E. coli* reactions as well as novel pathways that can serve as *de novo* biosynthesis pathways towards valuable chemicals.

5.1 Introduction

Genome-scale metabolic (GEM) reconstructions capture the known metabolic capabilities of organisms, and they have been used in many metabolic engineering studies that have provided important insights into complex biochemical networks. However, these studies rely on known biochemical reactions, and they may fail to exploit full metabolic potential of studied and engineered microorganisms as they neglect novel metabolic reactions, i.e. the reactions that are currently unknown to exist in the studied microorganisms.

The metabolic network of *Escherichia coli* is one of the best-characterized reaction networks in biology and several GEMs for different strains with different sizes have been developed for this microorganism since 2000 [190]. Nevertheless, even for well-studied organisms such as *E. coli,* there are still significant knowledge gaps and missing biotransformation steps to be characterized [170].

In the last decade, tremendous efforts were invested on developing systematic methods for the prediction and *de novo* design of novel pathways [99]. Most of the studies have focused on retrobiosynthesis tools to identify *de novo* pathways for the production of target molecules in a specific organism [74-77,82,191]. However, these strategies explore novel pathways only around the chemistry of the target structure and they neglect the potential evolution of metabolism.

However, applying these methods with a new perspective can help us to investigate a more fundamental questions: what is the utmost metabolic capacity of an organism such as *E. coli*,, and what are the metabolic functionalities that allow us achieving this? The answer to this question will provide a full list of all possible metabolites and metabolic reactions in the organism.

In this study, we introduce a systematic computational workflow for exploring the metabolic capacity of a given microorganism. Our approach combines the knowledge from genome scale models with a rule-based network reconstruction algorithm and results in a new generation of model organisms, so called “super organisms” with expanded metabolic capabilities.

The workflow has three main components (Figure 5.1):

(i) We employ the BNICE.ch framework, which is able to generate every possible biochemical reaction from a given set of generalized enzyme reaction rules and a set of starting compounds; BNICE generates a super metabolic network that explores the metabolic capability of the organism by introducing many novel reactions.

(ii) In order to identify the functionality of these novel reactions and to examine their feasibility for implementation purposes, we embed the novel reactions into the genome scale model of *E. coli* and we use a computational pipeline (composed of FBA, TFBA, and MILP optimization) to identify sets of reactions that can increase a desired metabolic property such as the biomass yield on glucose.

(iii) We further investigate the structural similarity of the novel reactions to the known enzymatic reactions. We used this similarity metric for the identification of candidate gene sequences for the novel reactions.

Starting from the known biochemistry of *E. coli* metabolism, we investigated how many novel reactions can be identified and if these reactions can give extra metabolic potential to this organism for industrial application in the production of fuels and chemicals. Interestingly, we found that there exist reactions in the generated super metabolic network that can increase the maximum theoretical biomass yield on glucose. This work allows for the first time the full exploration of theoretical metabolic capabilities of an organism through the addition of metabolic functionalities.

5.2 Methods

The proposed workflow consists of several computational steps (Figure 5.1), and the details of each step will be discussed in the following sections. Though we apply the workflow to construct the expanded metabolic network of *E. coli,* it can be used for any other organism.

5.2.1 Construction of a "super" metabolic network

The reconstruction of a "super" metabolic network of *E. coli* is accomplished using BNICE.ch computational framework [72].

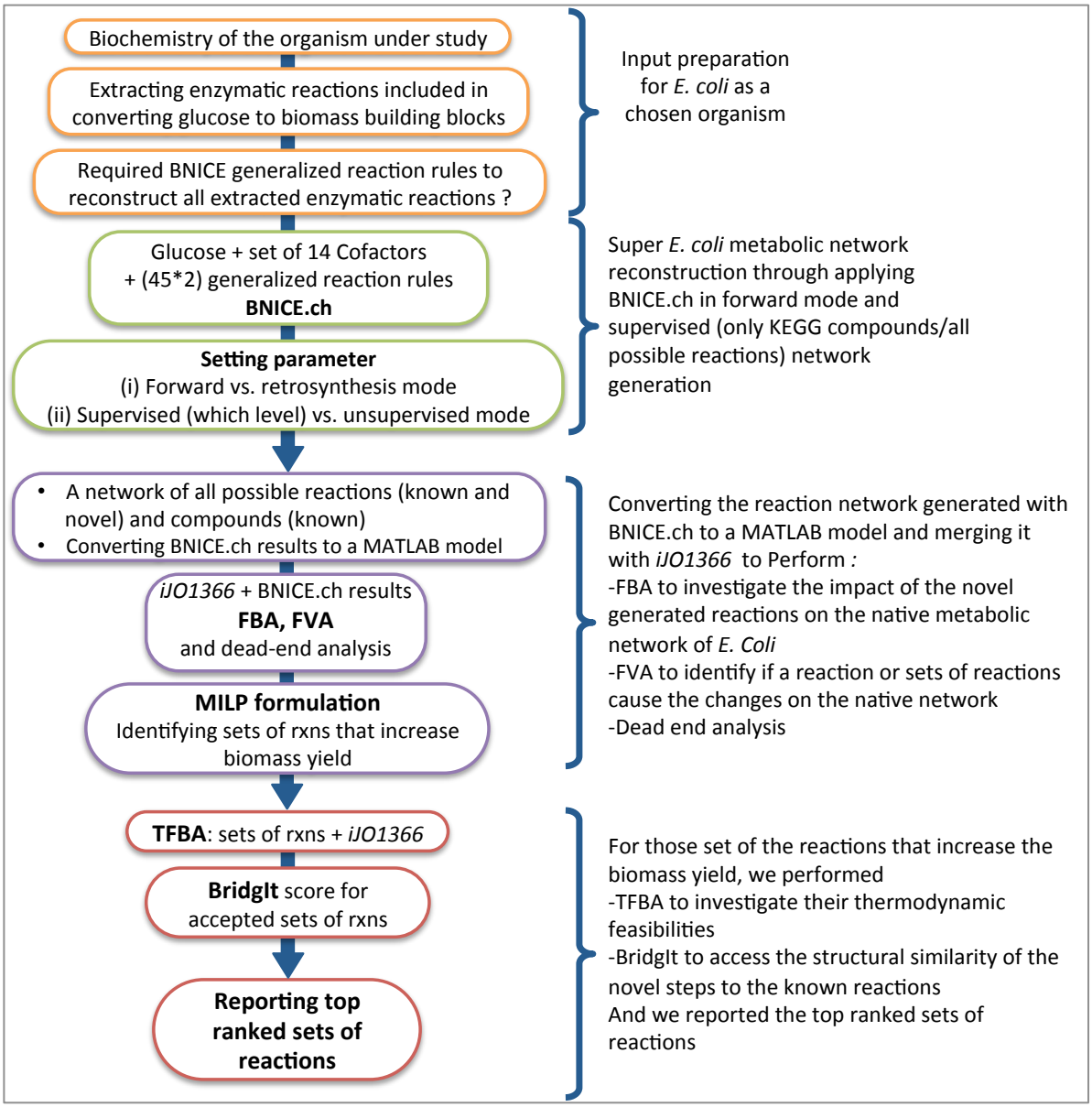


Figure 5-1 : The workflow for generating the "super *E. coli*" metabolic network that contains thousands of non-native enzymatic reactions (known and novel) and evaluating the impact of the novel steps to the enzymatic profile and metabolic capacity of this organism.

Different modules and methodologies of BNICE.ch has been discussed in Chapter 2, here we summarize the operating mode and the parameters that we chose in this study.

We focused on the central carbon metabolism since the breakdown of nutrients and the major part of carbon modification and energy metabolism is carried out in the core carbon metabolism. In addition, most of the metabolite targets for the metabolic engineering studies belong to the core carbon metabolism. Therefore, in our analysis we considered 5 central metabolism subsystems glycolysis, pentose phosphate pathway, citric acid cycle, glyoxylate cycle and pyruvate metabolism, together with the biosynthesis pathways for glycerol and the amino acids close the central carbon metabolism: glutamate, glutamine, aspartate, asparagine, alanine and serine.

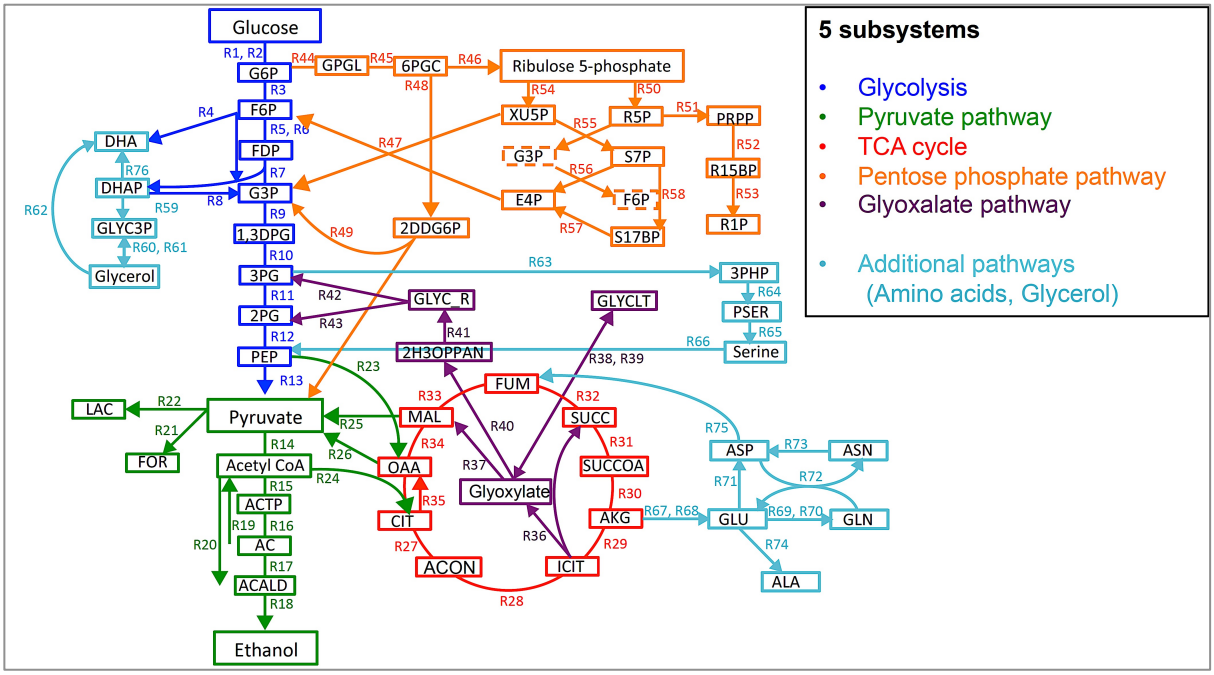


Figure 5-2 : The core metabolic network of *E. coli* contains 67 compounds including 14 cofactors. The compounds are connected by 76 enzymatic reactions which are attributed to one of the subsystems: Glycolysis (blue), Pyruvate pathway (green), TCA cycle (red), Pentose phosphate pathway (orange), Glyoxylate pathway (violet), or additional pathways accounting for the biosynthesis of some amino acids or glycerol (light blue).

We used the latest genome-scale metabolic model i*JO1366* [170,192] for *E. coli K-12* as a reference for extraction of known core enzymatic reactions discovered so far for this organism. The resulting network included 67 compounds (including 14 cofactors) and 76 reactions (Figure 5.2).

We needed 45 generalized reaction rules to reproduce these 76 known enzymatic

reactions using BNICE.ch. Since the directional feasibility of the reactions is further assessed through the TFBA analysis, we included the enzymatic reaction rules for both reaction directionalities (forward and reverse). Therefore, we applied 90 (45*2) reaction rules to reconstruct the core carbon metabolism of *E. coli*.

We used the "*forward network reconstruction mode*" of BNICE.ch, we left unconstrained the "number of iteration" for the system, and we allowed the network generation algorithm to run till the number of compounds and reactions converge, i.e. till no more compounds and reactions could be generated.

We were interested to investigate how many hypothetical reactions could evolve from the known biochemistry of *E. coli*. Therefore, we allowed only biologically known compounds to be selected as a part of the network, and we did not put any constraint on the reactions, which resulted on the generation of thousands of novel reactions in the reconstructed metabolic network of this organism. Table 5.1 shows the level of supervision we applied in this study.

Table 5-1: The supervision level for the reactions rules is on the network level, for the compounds on the database (KEGG) level and there was no constraint for the reactions and all the possible reactions (known and novel) are allowed in the network generation process

Supervision level		Input constraint		
	Example	Reaction rules	Compounds	Reactions
Unsupervised	-		Unlimited	**Unlimited**
Database	**KEGG**		**14549**	
Organism				
Network	**Core metabolic network of *E. coli***	**45*2**	67	76
Pathway				

5.2.2 Integration of novel reactions in genome scale model

To further analyze the thousands of hypothetical enzymatic reactions generated by BNICE.ch, we developed tools and frameworks to study their impact on the metabolism of *E. coli*. The genome scale metabolic reconstruction of *E. coli iJO1366* [192] was utilized as the reference model for integrating the novel generated reactions in the context of metabolic model, and further for performing *in silico* analyses such as Flux

Balance Analysis (FBA), yield calculations and gap filling.

E. coli genome scale network includes 2585 reactions and 1807 metabolites across cytoplasm, periplasmic space and growth media, and can grow on different carbon sources under aerobic and anaerobic conditions [192]. Since all the generated novel reactions are around the central carbon metabolism, we integrated them into the *E. coli* metabolic network as cytoplasmic enzymatic reactions. We call the genome scale model with the integrated new reactions the "*de novo E. coli* metabolic network" and we will use this term in the next sections.

5.2.3 Identifying sets of reactions that increase the biomass yield in *E. coli* network

We performed Flux Balance Analysis (FBA) to investigate if the novel proposed reactions increase the yield of biomass production. By performing a Flux Variability Analysis (FVA) on the extended metabolic network with the integrated *de novo* reactions, we observed that there was not a unique set of reactions that would increase the biomass yield. Hence, we developed an iterative method for identifying sets of reaction(s) that increase the yield, and we have determined all possible combinations of novel reactions responsible for this increase. We outline the developed method as follows:

i) Split the reactions in *de novo* network into forward and reverse reactions

ii) Add binary use variables for each of the novel reactions

iii) Formulate the problem as:

$$\text{Max} \sum z_i$$

s.t.

$$S.v = 0$$

$$v_{min} \leq v_i \leq v_{max}$$

$$v_{f,i} + v_{b,i} + Kz_i \leq K$$

z_i – binary variable associated with the reaction *i*

S – stoichiometric matrix of the network

v – flux vector and v_{min} , lower and v_{max} upper bound for the reaction fluxes

K – large arbitrary constant to enforce the thermodynamic constraint

We have also included constraints to prevent the simultaneous usage of forward and reverse reactions; hence either forward or reverse direction can be active.

iv) However, this formulation will generate only one solution, and we have already determined there is more than 1 set of reactions that increases the yield. To enumerate all possible sets, we have created a constraint that prevents the solver from choosing the same solution, and we have exhaustively enumerated all possible sets.

5.2.4 Thermodynamic-Based Flux Balance Analysis (TFBA)

We did not have any experimental evidence about the kinetic (ir)reversibility of the possible biotransformations, and thus we integrated novel reactions into the native *E. coli* network as bidirectional reactions. Since thermodynamic properties of reactions affect their directionalities, we tested the feasibility of the sets of reactions integrated in the *E. coli* genome scale network under thermodynamic constraints. For this purpose, we utilized the formulation of Thermodynamics-based Metabolic Flux Analysis (TMFA/TFBA) [47], which uses Group Contribution Method (GCM) [27] to estimate the standard Gibbs free energy of formation of metabolites, and builds a Mixed Integer Linear Programming (MILP) problem to incorporate the thermodynamic constraints in metabolic networks. We applied these constraints [48,193,194], and we verified the thermodynamic feasibility of the proposed novel reaction sets and their impact on the yield of biomass accumulation.

5.2.5 Gap-filling analysis by using novel reactions in *E. coli* network

There are knowledge gaps in the latest proposed genome scale reconstruction (GEM) of *E. coli* [170]. This is due to the missing enzymatic functions, which were not captured during gene to reaction association (GPRs) and the gap filling analysis. We have utilized BNICE.ch generated reactions to identify possible gap filling reactions for the dead-end (blocked) metabolites in *E. coli* metabolic network. We performed the dead-end analysis of the *de novo* generated metabolic network, and we reported the metabolites that were dead-ends in the native network but connected to the metabolism in the enlarged network.

5.2.6 BridgIt analysis

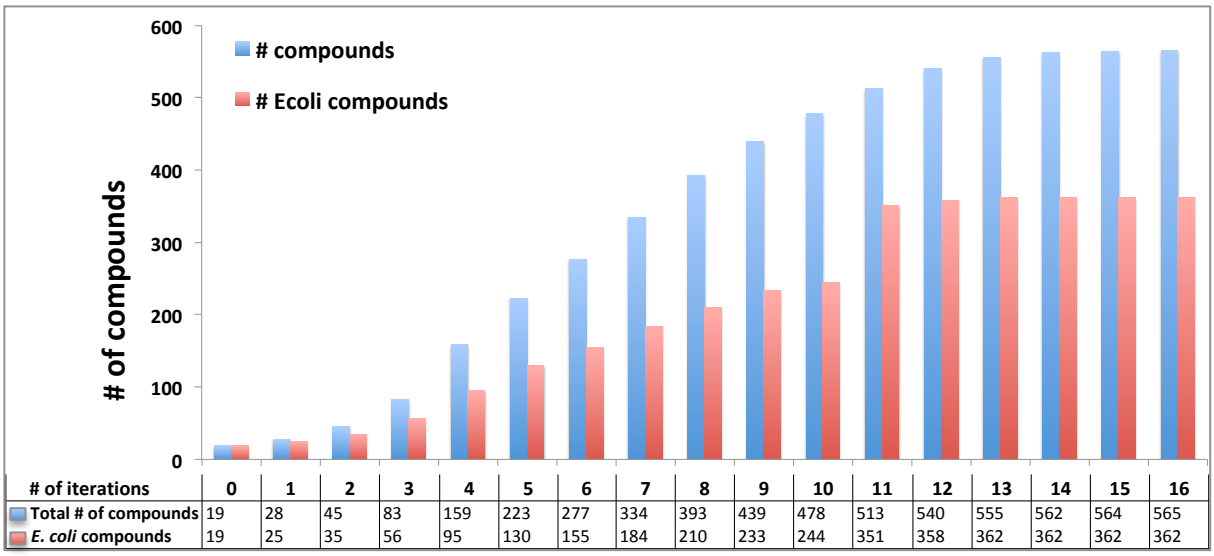
In Chapter 2 we described the methodologies behind the BridgIt framework and in Chapter 3 we demonstrated its efficiency on finding the structurally similar known reactions to the hypothetical generated reactions of BNICE.ch.

5.3 Results and discussions

We explored how the metabolic network of *E.coli* evolves within the known biochemistry if we apply the aforementioned 90 generalized enzymatic reaction rules to glucose as a starting substrate along with the set of relevant cofactors. We also analyzed how the generated hypothetical reactions affect the known metabolic network of *E. coli* and how these reactions are changing the enzymatic profile of this network.

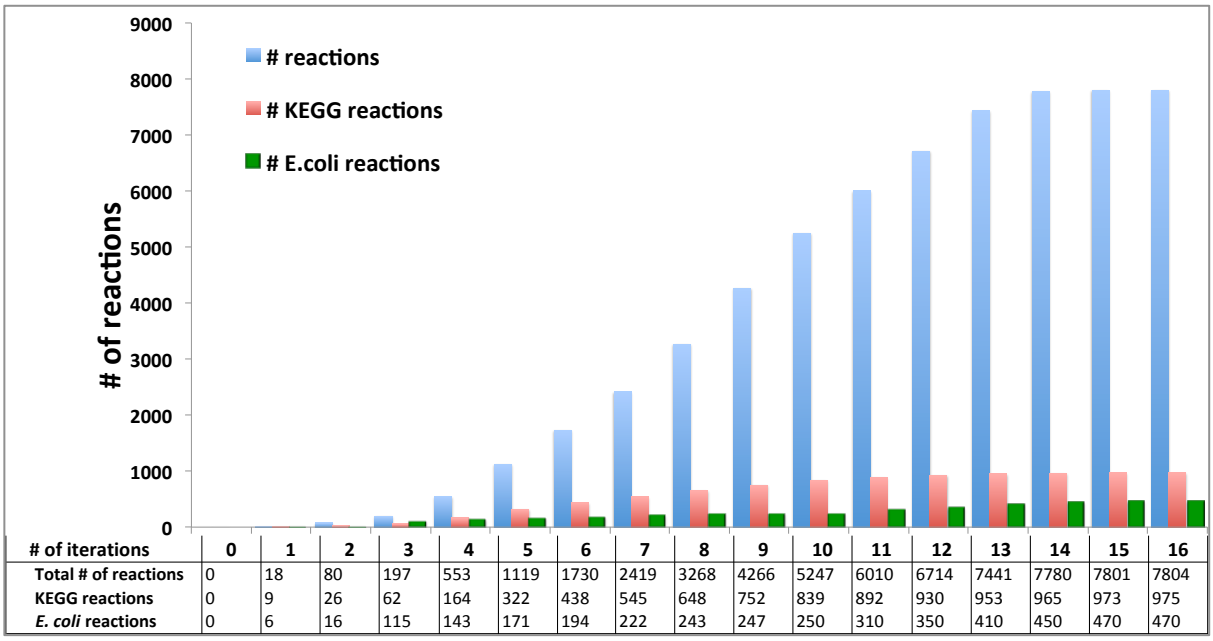
5.3.1 *E. coli* super metabolic network reconstruction

The network generation algorithm converged after 16 iterations resulting in a network of 565 compounds and 7804 reactions. 975 out of 7804 generated reactions were already reported in KEGG database (Figures 5.3 & 5.4). The algorithm reconstructed successfully 67 compounds and 76 reactions that exist in the core metabolic network of *E. coli* (see Section 5.2.1), and even more striking was that we obtained additional native *E. coli* compounds and reactions (Figures 5.3 & 5.4). This demonstrated that the utilized generalized reaction rules could be used to explain the biochemistry of the other parts of *E. coli* metabolism.



# of iterations	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Total # of compounds	19	28	45	83	159	223	277	334	393	439	478	513	540	555	562	564	565
E. coli compounds	19	25	35	56	95	130	155	184	210	233	244	351	358	362	362	362	362

Figure 5-3: Total number of generated compounds (all KEGG compounds) after 16 iterations along with the numbers of generated native *E.coli* compounds. After 16 iterations, the number of compounds converged to 565.



# of iterations	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Total # of reactions	0	18	80	197	553	1119	1730	2419	3268	4266	5247	6010	6714	7441	7780	7801	7804
KEGG reactions	0	9	26	62	164	322	438	545	648	752	839	892	930	953	965	973	975
E. coli reactions	0	6	16	115	143	171	194	222	243	247	250	310	350	410	450	470	470

Figure 5-4: Total number of reactions, KEGG reactions and *E.coli* native reactions generated in each iteration through the reconstruction of core metabolism of *E.coli*. The total number of reactions converged to 7804 after 16 iterations.

5.3.2 Characteristics of *de novo* metabolic network of *E. coli*

After integrating the generated compounds and reactions into the genome scale model of *E. coli*, the overall number of metabolites in the extended network increased by

20.4% to 2176 compounds, whereas the number of reactions (7804) almost tripled compared to the native model. 362 out of 565 generated metabolites and 470 out of the 7804 generated reactions were already native for *E. coli*.

Since reactions in the metabolic models are considered as bidirectional unless the kinetic irreversibility is observed experimentally, we added to the expanded model the forward and reverse reactions of the same BNICE.ch biotransformation as one bidirectional reaction. We ended up with the expanded model with 6454 reactions and 2176 metabolites.

One of our main goals in this study was to investigate the possible contributions of the proposed novel reactions to the *E. coli* metabolism and to determine if these reactions could be used to fill in some missing functions in the current functional annotation. Firstly, we tested the *de novo* metabolic network for its capacity to accumulate biomass precursors. Wild type *E. coli* grows with a specific growth rate of 0.99 1/hr with a glucose uptake of 10mmol/gDWhr under aerobic conditions.

To assess the impact of the novel introduced hypothetical reactions to the metabolic network of E. coli, we performed Flux balance Analysis (FBA) on the *de novo E. coli* network and we obtained a specific growth rate of 1.46 1/hr under the same media composition, which corresponds to 46% increase in the biomass yield compared to the wild type *E. coli*.

5.3.3 Sets of reactions that increase the biomass yield

We performed a Flux Variability analysis (FVA) on the *de novo* metabolic network, and we found out that there were no individual novel reaction(s) whose addition to the *E. coli* network would increase the biomass yield on glucose. This implied that there were alternative set(s) of reactions that were responsible for this increase in the yield. In order to determine these different sets of reactions, we applied our framework (See Section 2.2.1), and we identified all possible sets that increase the yield in the *de novo* network (Table 5.2).

Table 5-2: Total number of generated sets of reactions that increase the yield in FBA, and the number of those that are network thermodynamically feasible and that increase the yield in the TFBA analysis as well.

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

We found that 11048 different sets of reactions with lengths 2, 3, 4 and 5 could increase the yield to at least 90% of the theoretical optimum yield (1.3 1/hr specific growth rate with 10 mmol/gDWhr glucose uptake rate). These 11046 sets were composed of only 1140 reactions (~15 percent of the total), and 135 of them were known KEGG reactions whereas the others were novel reactions proposed by BNICE.ch.

5.3.4 TFBA of *E. coli* metabolic network along with different sets of reactions

After populating all possible combinations of the novel reactions that increase the yield, we have tested them under thermodynamic constraints to investigate if the bioenergetics would allow those biotransformations to operate in the favorable direction.

We integrated the identified sets one at a time into the *E. coli* native metabolic network, and we have performed the TFBA analysis of so obtained extended metabolic networks. The highest specific growth we obtained with these extended networks was 1.39 1/hr under aerobic conditions, with 10 mmol/gDWhr glucose uptake. The first obvious result was that due to the thermodynamic constraints the highest specific growth rate decreased from 1.46 1/hr to 1.39 1/hr. Even though some identified sets of reactions were thermodynamically feasible, some reactions that belonged to the native metabolism were not thermodynamically feasible in the direction that was imposed by the integrated sets of novel reactions. As a result, we observed a drop in the yield compared to FBA results. On the other hand, 8285 identified sets were thermodynamically infeasible in the increasing yield direction (Table 5.2)..

During the analysis of some sets, we could not conclude the main reason for the biomass yield increase at a first glance, and we performed further analysis to understand the behavior of the *de novo* network.

We built *de novo* models with the sets that increase the yield to the highest with thermodynamic constraints, and performed TFBA simulations to observe the behavior of the *de novo* network. By analyzing these models, we concluded that the yield increase is mostly related to ATP, which states that the limiting metabolite for the biomass yield is ATP, rather than any other biomass building block. The detailed study of these sets is an ongoing project and it is subject to further analysis.

5.3.5 Analysis of dead-end metabolites

In the de novo metabolic network of *E.coli*, we had a closer look to the metabolites that are "dead end metabolites" in the native model of *E. coli* and we investigated their connectivity in the new generated network in attempt to identify the new reaction steps that connect them to the other metabolites in the network.

In total, there were 208 dead-end metabolites in *i*JO1366 which represent the knowledge gaps in this model reconstruction [170]. Nine metabolites that were dead-end in the native model got connected with the other metabolites of the *de novo* metabolic network through the novel reactions of BNICE.ch (Table 5.3).

Table 5-3: 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 that are not dead-end in the *de novo* metabolic network	**# of reactions could connect the metabolites to the native metabolites**	**# 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

5.3.6 BridgIt analysis

Since in this analysis we explored all the possible enzymatic reactions of *E. coli* on the basis of its known biochemistry (the network includes only known compounds), the set of generated novel reactions were a subset of the hypothetical reactions we reported in Chapter 3. Therefore their detailed analysis is discussed in Chapter 3. Hereby, in order to compare different sets of reactions and to rank them based on their structural similarities to the known reactions, we reported an "average BridgIt score" for the reactions in each set.

5.4 Conclusions

E. coli is one of the most studied organism in the metabolic engineering field and many research works have proven its capacity for the biosynthesis of high value chemicals and its flexibility to metabolic modifications.

In order to make use of the full production potential of *E. coli*, in this study we attempted to reveal its full enzymatic capability based on the generalized enzymatic reaction rules of BNICE.ch. We used our method to generate a "super" metabolic network that included: (i) all the metabolites and reactions of the native core metabolism of *E. coli*; (ii) several known but non-native metabolites and reactions (iii) and hundreds of novel reactions that linked known and non-native metabolites. The network was further evaluated with respect to the thermodynamic feasibility of the novel generated reactions in the context of the genome scale model of *E. coli iJO1366.*

We identified several sets of reactions that increase the yield through biomass production, and that could also be implemented as *de novo* pathways for the biosynthesis of several heterologous compounds.

Chapter 6

iAM.NICE!

What is the fate of each atom from nutrients into metabolic universe?

SUMMARY

In this chapter, we introduce a computational framework, "iAM.NICE" (*in silico* Atom Mapped Network Integrated Computational Explorer), for the atom-level reconstruction of metabolic networks from the *in silico* labeled substrates, which allows tracking the fate of atoms through the reconstructed metabolic network. The originality of "iAM.NICE" is twofold. First, it is to our knowledge the first automated atom-mapping algorithm that is derived from the underlying enzymatic biotransformation mechanism; and second, its application is not limited to individual reactions and it can be used for the reconstruction of atom-mapped metabolic networks. We illustrate the effectiveness of our method through the reconstruction of atom-mapped reactions of KEGG database. Furthermore, we provide an example of an atom-level representation of core metabolic network of *E. coli* to show how a comprehensive atom level metabolic analysis can guide the experimental design to obtain more precise biological information.

6.1 Introduction

The automated identification of atom transitions within a reaction pathway is a very challenging task since the degree of complexity of metabolic networks dramatically increases when we transit from *metabolite-level* studies to *atom-level* studies. Despite being studied extensively in various approaches, the field of atom mapping of metabolic networks is lacking an automated approach, which accounts for the information of *reaction mechanism* for atom mapping and is extendable from *atom-mapped reactions* to *atom-mapped reaction networks*.

In a reliable *in silico* atom-level reconstruction of a metabolic network the atom mappings need to be: (i) correctly created at the level of individual reactions, and (ii) connected in a network and conserved through all reactions steps from the input compound to the final products, enabling one to trace back the exact metabolic path of every single atom. To do so, we developed the "iAM.NICE" framework for addressing the atom-mapping problem.

In this study, we provide the atom-mapped representation of all the reactions in KEGG database as well as an atom-level representation of core *E-coli* metabolic network. Our results can be used for a large range of applications, starting from the identification of new metabolic routes for the microbial production of desired compounds to the interpretation of the optimum labelled patterns of substrates, which can be a great benefit for the simulation of tracer experiments.

6.2 Methods

6.2.1 In *silico* atom labelling

In the "iAM.NICE" framework, molecules are represented in the form of a Bond-Electron-Matrix (BEM) that describes the electron bonds between the atoms [101] and we store this information as a molfile [195]. Our method allows us to automatically label all the atoms of a molecule, with the exception of hydrogen atoms. In our *in silico* labelling studies we adopted the annotation used in stable isotope ^{13}C labelling experiments. For the cases where we needed to trace the fate of more than one labelled atom, we used ^{14}C, ^{15}C, etc. Though all the examples shown in this work represent

carbon labelling, the same principle can be applied for the annotation of other atom types.

Within the "iAM.NICE" framework, we can label substrate molecules in many distinctive ways by altering the number of labelled atoms and their positions. However, in this study we will focus on the following cases:

1) Fully labelled molecules: fully labelled molecule means that all labelled atoms in a molecule have distinct labels, whereas in stable isotope ^{13}C labelling experiments all labelled atoms have the same label. We use such labelled molecules to construct atom-maps of individual reactions; here, the atoms of the starting compound(s) are labelled, with the exception of hydrogen atoms, and the result of their transition through a biochemical transformation can be observed in the atom-mapped reaction product.
2) Single atom labelled molecules: we use these molecules to construct atom-mapped pathways and metabolic reaction networks; here, a single atom of the starting compound is labelled, and the fate of this atom can be observed through a sequence of reactions.

Figure 6.1 shows an example of *in silico* carbon-labelled glucose and the comparison between different isotopomers of glucose. The term isotopomer comes from isotope isomers, and it describes different isotopic compositions of the same compound. An important characteristic of "iAM.NICE" is that it considers different isotopomers as different compounds, enabling the differentiation between isotopomers and hence the reconstruction of atom-mapped networks.

Depending on the purpose of the *in silico* labeling study, we choose a different isotopomer as starting compound. Unlabeled glucose is compared with fully carbon labeled and different C-labeled glucose based on the position of the labeled carbon. If we want to follow a specific atom of the starting compound through the network, we choose ^{13}C-single labeled glucose. C1 to C6 specify the position of the labeled atom in the molecule.

Note that we can also use a fully labelled molecule in sequences of reactions. However, with all the atoms being labelled, as the size of metabolic network increases the generated atom-mapped network becomes prohibitively complex for any practical

purpose. Thus, we use the single atom labelled molecules to study the metabolic networks.

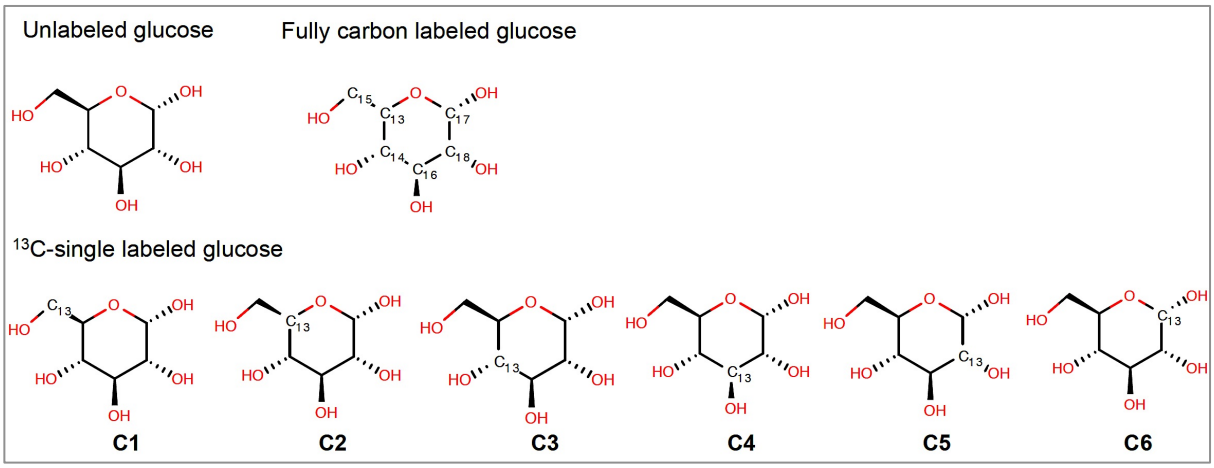


Figure 6-1: Depending on the purpose of the *in silico* labeling study, we choose a different isotopomer as starting compound. Unlabeled glucose is compared with fully carbon labeled and different C-labeled glucose based on the position of the labeled carbon. If we want to follow a specific atom of the starting compound through the network, we choose ^{13}C-single labeled glucose. C1 to C6 specify the position of the labeled atom in the molecule.

6.2.2 Atom-mapped reactions: transferring labelled substrates to labelled products

As already discussed in Chapter 3, the *generalized enzymatic reaction rules* were initially developed in BNICE.ch with the aim of discovering *de novo* alternative biosynthesis and biodegradation pathways, and their potential has been demonstrated in several studies [72,82,84].

Contrary to their application in our previous studies, in this study, for the first time we applied the generalized reaction rules not for generating *novel* information, but for introducing another level of information to the *existing knowledge* of metabolism. Specifically, we added detailed information about the atom transition to the *known* metabolic networks by reconstructing the *known* enzymatic reactions at the atomic level. For this purpose, we adjusted the BNICE algorithms, and we also formulated new algorithms for the atom-mapped reconstruction of enzymatic reactions, pathways and metabolic network.

Our algorithm read the KEGG reactions, one by one, extracted the substrates and checked them against a list of cofactors given as a parameter and after identifying the cofactors, it automatically regenerated the *in silico* labelled non-cofactors substrate by

successively labelling their carbon atoms. In case of having more than one substrate, the successive manner continues after labelling the first substrate (Figure 7.3, EC Class2).
In order to generate atom-mapped reactions, "iAM.NICE" transfers the label(s) from a substrate to a product by taking into account the information about the rearrangements of the chemical bonds derived from the generalized reaction rules. This way, we generate the product structure from the substrate structure by rearranging bonds, while conserving the information about the positions of atoms. Figure 6.2 shows an example of atom-mapped reaction along with the BEM of representation of substrate, products and the matrix representation of the corresponding generalize reaction rule for carrying out such an enzymatic reaction. The colour codes help to clarify the different corresponding parts of BEM and reaction rule matrix. According to the mechanism predefined in the reaction rule, the $^{13}C-{}^{14}C$ bond breaks and two products form. The cofactor of the reaction (NADPH) rearranges the electrons as described in the reaction rule to generate a double bond instead of $^{13}C-O$ and $^{15}C-O$ which results on the formation of labelled CO_2 and pyruvate as the products with specified carbon position being conserved based on the mechanism of the reaction.

6.2.3 Atom-mapped pathways and networks

In our proposed approach, the atom-mapped pathways and networks are reconstructed in an iterative manner. In the first iteration of the procedure, we apply the generalized reaction rules on a starting *in silico* labelled substrate. If one of the generalized reaction rules recognizes the reactive site in the labelled substrate, the reaction rule rearranges the bonds according to its distinct reaction mechanism. This way, we obtain a set of atom-mapped reactions together with their labelled products. The labelled products obtained in the first iteration are the substrates for the second iteration. As a result, the labelled products of the second iteration are two steps away from the initial labelled substrates and they contain the information about the fate of each atom after these two reaction steps.

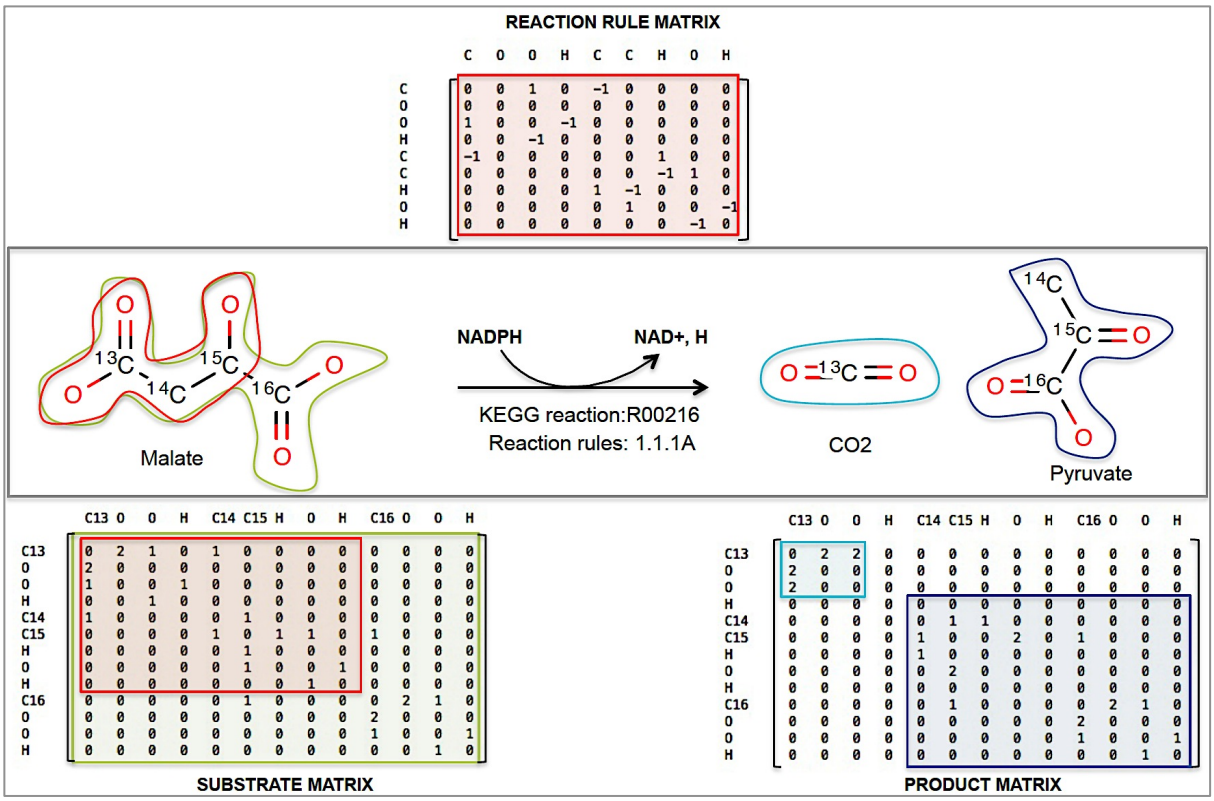


Figure 6-2: BEM representation of an atom-mapped example reaction: The generalized reaction rule recognizes the reactive site (red) of the substrate. Next, the reaction rule matrix, which stores the bond changes for every pair of atoms, is added to the substrate matrix (green) at the reactive site. The resulting product matrix contains the BEM of the products: CO_2 (light blue) and Pyruvate (dark blue).

We iterate the steps of the procedure till we reconstruct studied pathways or networks. This way, we can trace the fate of every single atom from the first substrate to the final product, i.e. we eventually obtain a comprehensive atom level reconstruction of metabolic pathways or metabolic networks.

An important feature of "iAM.NICE" is that the existing metabolic data in biological databases is integrated into our framework and is organized within different categories and levels such as compounds, reactions, pathways, organisms, etc. This allows us to screen the obtained results against all known biological knowledge in different levels. Detailed explanations on the different levels of supervision and their importance for the "i.AM.NICE" framework are provided in Chapter 2.

6.3 Results and discussion

In the following we present a proof of concept for the atom mapping of enzymatic reactions, pathways and metabolic networks using "iAM.NICE", we continuously elaborate on the computational and methodological concepts associated with this approach, and we illustrate the application of the proposed method for the atom mapping of the *E.coli* central carbon metabolism.

6.3.1 Atom mapping for single enzymatic reactions

We used our framework to reconstructed atom-mapped KEGG reactions (version July 2014). We used fully carbon labeled KEGG molecules (Section 6.2.1) as inputs to the algorithm together with all the generalize reaction rules. Upon acting on a molecule, the generalized reaction rules automatically identify the biological reactive sides of molecule and apply the biotransformation by which the atoms and bonds rearrange to form the product. We show an example of a carbon mapped reaction for every enzymatic class of the Enzyme Commission (EC) classification system[196] in Figure 6.3.

6.3.2 Atom mapping for metabolic pathways

Biological pathways are a combination of several enzymatic reactions occurring in a sequence. "iAM.NICE" not only accurately maps atoms for individual reactions, but also extends these atom-mapped reactions to atom-mapped metabolic pathways and eventually the metabolic networks. In the following sections we presented the properties and utility of our approach by studying the central carbon pathways. Central carbon metabolism takes different enzymatic steps to convert sugars into metabolic precursors, which are used downstream to form the entire biomass of the cell. While in this study we focus on mapping carbon atoms, our method can be used to map atoms of all elements with the exception of Hydrogen.

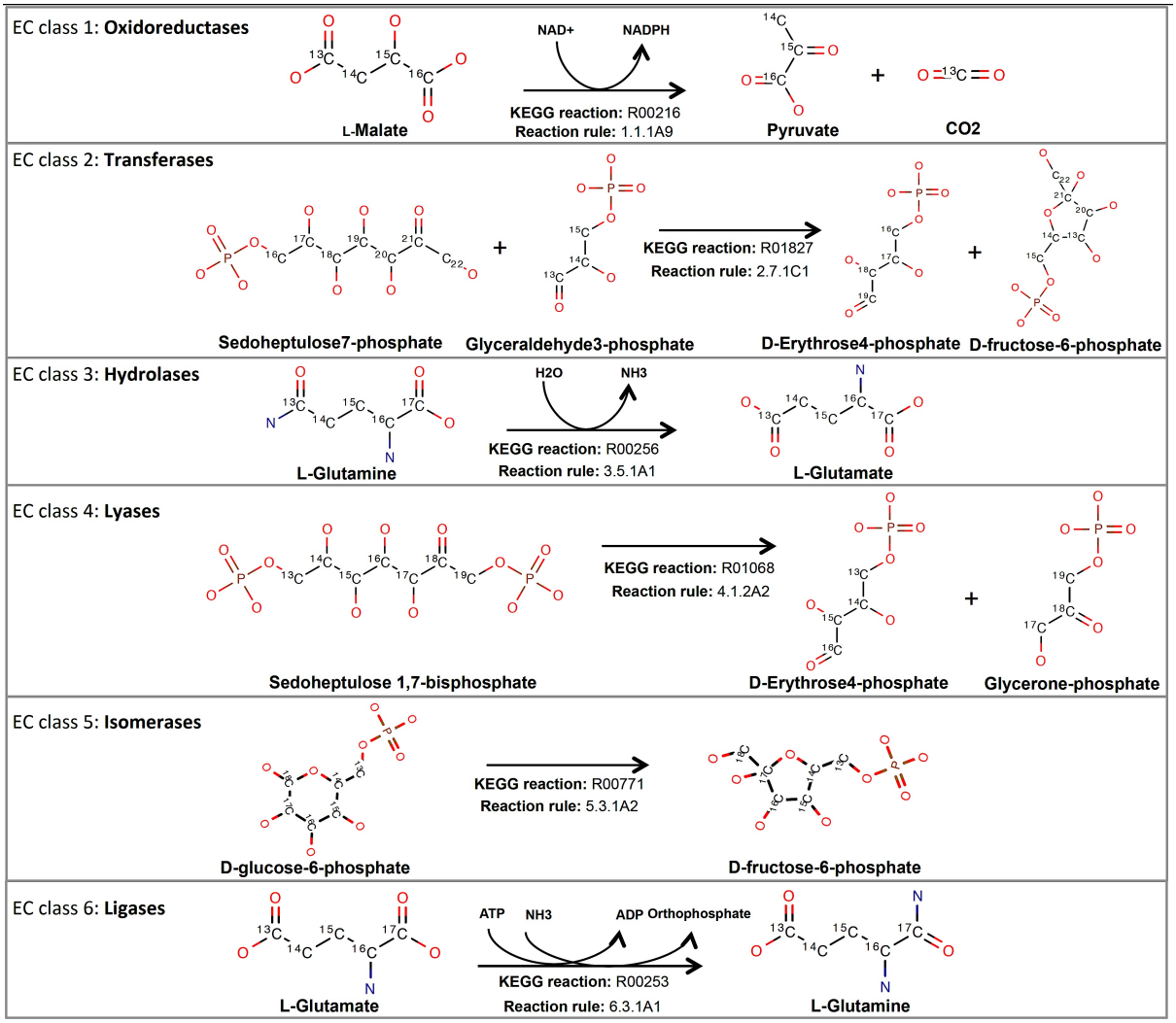


Figure 6-3: An example of a carbon atom-mapped reaction is given for each EC class. Carbon atoms in the substrates are labelled from ^{13}C upwards, and every carbon isotope present on the reactant side reappears on the product side. In the case of more than one substrate, e.g. the example of EC class 2, the carbon labels are enumerated sequentially for all the substrates. The molecular structures of cofactors that do not participate in carbon transfer are not shown. We show below the reaction arrows the KEGG Id and the name of the catalysing reaction rule.

6.3.2.1 Atom-mapped Glycolysis

Glycolysis is a metabolic pathway that converts glucose to acetyl-CoA and with a few variations, occurs in roughly all organisms. To generate the atom-mapped glycolysis, we constrained the scope of the search space of compounds and reactions to the pathway level (Section 7.2.3, Table 7.1) and we performed 10 iterations since the glycolysis pathway is a sequence of 10 enzymatic reaction steps, one should be careful that in some steps more than one reaction occurs which makes in total 16 reactions in

glycolysis (Figure 6.4). Since the entry point to glycolysis starts is D-glucose, we used as inputs for our algorithm the fully carbon-labelled glucose and the set of 9 cofactors that are required to reconstruct all the reaction steps in glycolysis (Figure 6.4)

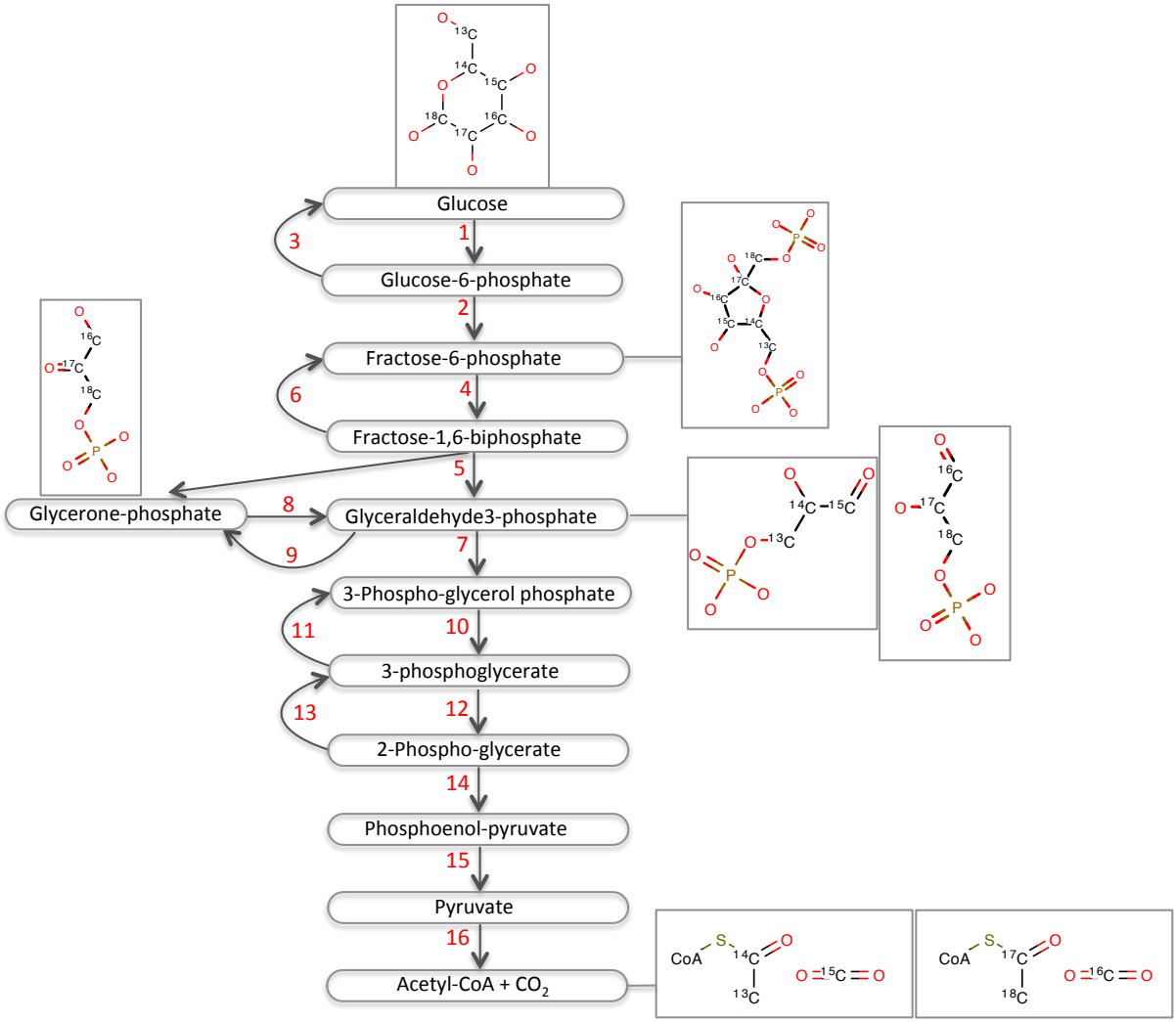


Figure 6-4: Glycolysis pathway reconstructed from a fully carbon-labelled glucose. Reactions are numbered from 1 to 16, connecting the 22 compounds and isotopomers. Boxes show the atom-mapped molecular structures of the produced isotopomers for a selection of metabolites.

In the first iteration, we applied the reaction rules on the labelled glucose, and we looked up in the sets of glycolysis reactions and compounds (predefined search space) the resulting atom-mapped reactions and labelled products. If both the product compounds and the reactions were part of the set of glycolysis compounds and the set of glycolytic reactions, respectively, the reactions were added to the pathway, and the product metabolites were added to the list of substrates for the following iteration. In

this example, the first metabolite generated from glucose was glucose-6-phosphate. In the second iteration, the generalized reaction rules were applied to glucose-6-phosphate, which resulted in the generation of fructose-6-phosphate. This procedure was continued until acetyl-CoA and CO_2 were produced and all the reactions in the list of glycolysis were reconstructed along with their atom maps (Figure 6.4). The resulting network had 22 compounds (including 9 pairs of isotopomers) connected by 16 reactions.

The isotopomer pairs resulted from the branching that occurred in reaction 5, where fructose-6-phosphate was split into glycerone-phosphate and glyceraldehyde3-phosphate, both carrying differently labelled carbon atoms. Glycerone-phosphate was further transformed into a second isotopomer of glyceraldehyde3-phosphate. The rest of the labelled pathway contained two parallel sequences of compounds, where each sequence followed a different labelling pattern.

We ended up with two isotopomers of acetyl-CoA and CO_2, with each molecule containing a subset of the initial isotope labels of the starting compound. Interestingly, the two carbon atoms giving CO_2 were initially present at the 3^{rd} and 4^{th} position in glucose, while the other atoms of glucose ended up in acetyl-CoA. This example demonstrated how we could trace back the itinerary of each single atom through the pathway. Other carbon-mapped biological pathways such as citrate cycle, pentose phosphate and pyruvate pathways can be replicated in the same manner and encompassed as a central carbon metabolism together and are presented in the next part.

6.3.3 Atom mapped *E. Coli* core metabolic network

In order to demonstrate the efficiency of "iAM.NICE" in reconstructing atom level metabolic networks, we mapped the carbon atoms of a core metabolic network of the *E.coli* core metabolism introduced in section 5.2.1, Figure 5.2. The network was a manually curated subset of the genome-scale metabolic model *iJO1366* for *E. coli* K-12 MG1655 [113], and it covered glycolysis/gluconeogenesis, pentose phosphate pathway, pyruvate metabolism, Tricarboxylic Citric Acid (TCA) cycle and glyoxylate metabolism, as well as the biosynthesis pathways for glycerol.

We constrained the scope of the search space of compounds and reactions to the network level, i.e. to the core metabolism of *E. coli* (Table 6.1), we used as inputs ^{13}C-single labelled glucose together with 14 cofactors involved in core *E. coli* metabolism, and we studied the flow of labelled carbon atom through the metabolic network. Starting from a glucose molecule we performed 16 iteration steps to cover every reaction by iterative network reconstruction.

Table 6-1: The level of supervision that we applied for the reconstruction of the atom-mapped core metabolic network of *E .coli* is at the network level for the generalized reaction rules, compounds and reactions

Supervision level		Input constraint		
	Example	Reaction rules	Compounds	Reactions
Unsupervised	-			
Database				
Organism				
Network	**Core metabolic network of *E. coli***	**45*2**	**67**	**76**
Pathway				

We repeated this procedure six times where in each repetition we changed the position of the labelled carbon atom in the ^{13}C-single labelled glucose molecule, i.e. we used as inputs carbon-labelled glucose C1, C2, C3, C4, C5 and C6 (Figure 6.1). This way, we generated six different atom-mapped networks.

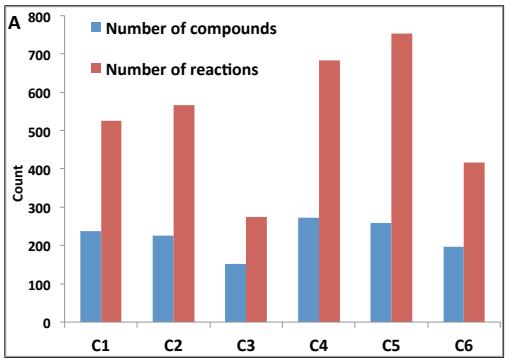
We analyzed the 6 resulting atom-mapped metabolic networks, and we observed that the size of networks (graphs), in terms of number of isotopomers (nodes of the graph) and number of reactions (edges of the graph), was changing depending on which labelled glucose molecule, C1 to C6, was used as the input. In fact, the position of the labelled atom in the labelled glucose molecule affected how the labelled atom was propagating through the network, and therefore how many isotopomers were generated. We identified four properties that gave rise to the differences in the size of the labelled metabolic networks:

I. Loss of ^{13}C to CO_2 during the course of a reaction reduces the number of isotopomers.

II. Production of labelled acetyl-CoA increases the number of isotopomers.

III. Presence of the label in the glycerol pathway increases the number of isotopomers.

IV. Labelled pyruvate originating from the pentose-phosphate pathway increases the number of isotopomers.



B

Statistics of generated compounds and reactions	unlabeled	Label position					
		C1	C2	C3	C4	C5	C6
Number of compounds	67	237	225	151	273	258	196
Number of reactions	100	526	567	274	684	754	417
Generations	16	16	16	16	16	16	16
^{13}C in Glycerol	-	no	no	no	yes	yes	yes
PP pathway gives labeled pyruvate	-	no	no	no	yes	yes	yes
label goes to Acetyl-CoA	-	yes	yes	no	yes	yes	yes
label lost to CO2	-	no	no	yes	yes	no	yes

Figure 6-5: Altering the position of the label (from C1 to C6) influences the size of the network. The barplot in Panel A shows numbers of compounds (blue) and reactions (red) for different positions of labelled carbon atom. The table in Panel B presents the number of isotopomers and reactions that are produced over 16 iterations, for each of 6 atom-mapped networks (i.e. for altered inputs from C1 to C6). Red cells (or alternatively *green cells*) in Panel B show that the corresponding property if satisfied reduces (or *increases*) the network size by promoting (or *impeding*) the generation of new isotopomers.

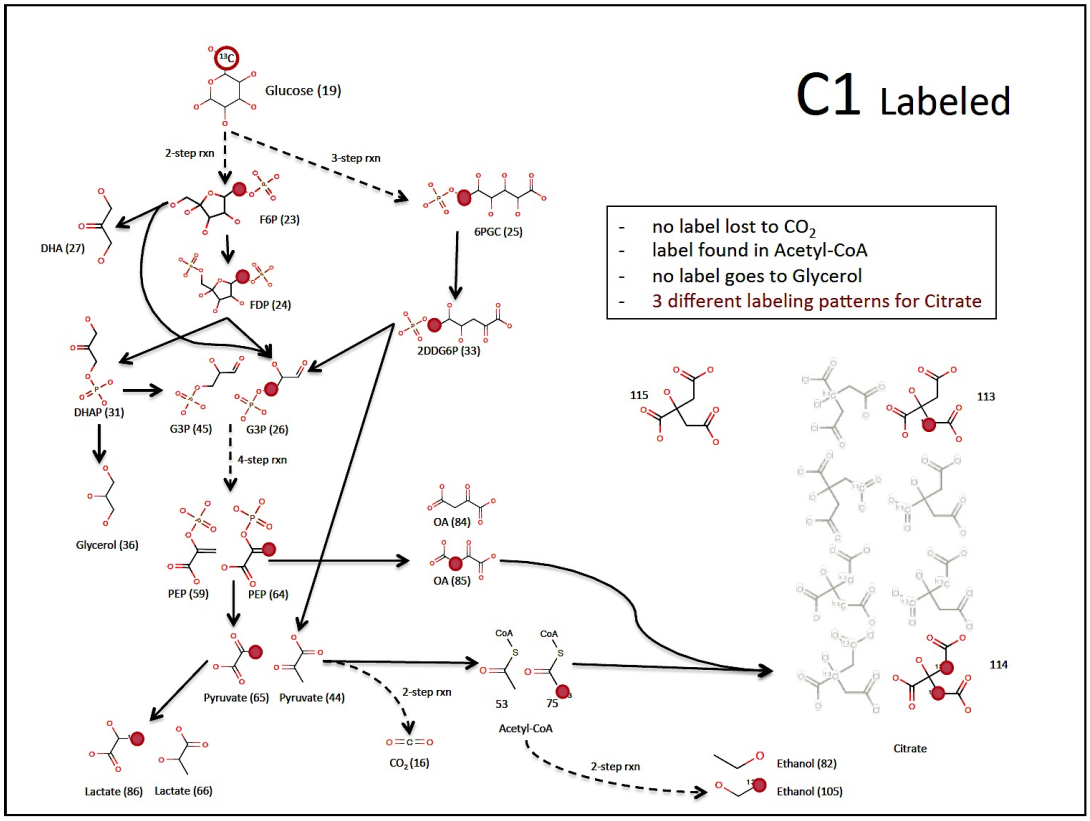
We compared the statistics of generated isotopomers and reactions in 6 generated networks (Figure 6.5, Panels A and B). Labelling glucose at position 3 resulted in a relatively small network (151 isotopomers). In contrast, labelling positions 4 and 5 resulted in the generation of 273 and 258 isotopomers, respectively. The network sizes of the remaining labelling positions lied in between these values. We next analyzed the occurrence of the above-mentioned properties I-IV (Figure 6.5, Panel B), and we evaluated whether or not these 4 properties were true for six networks.

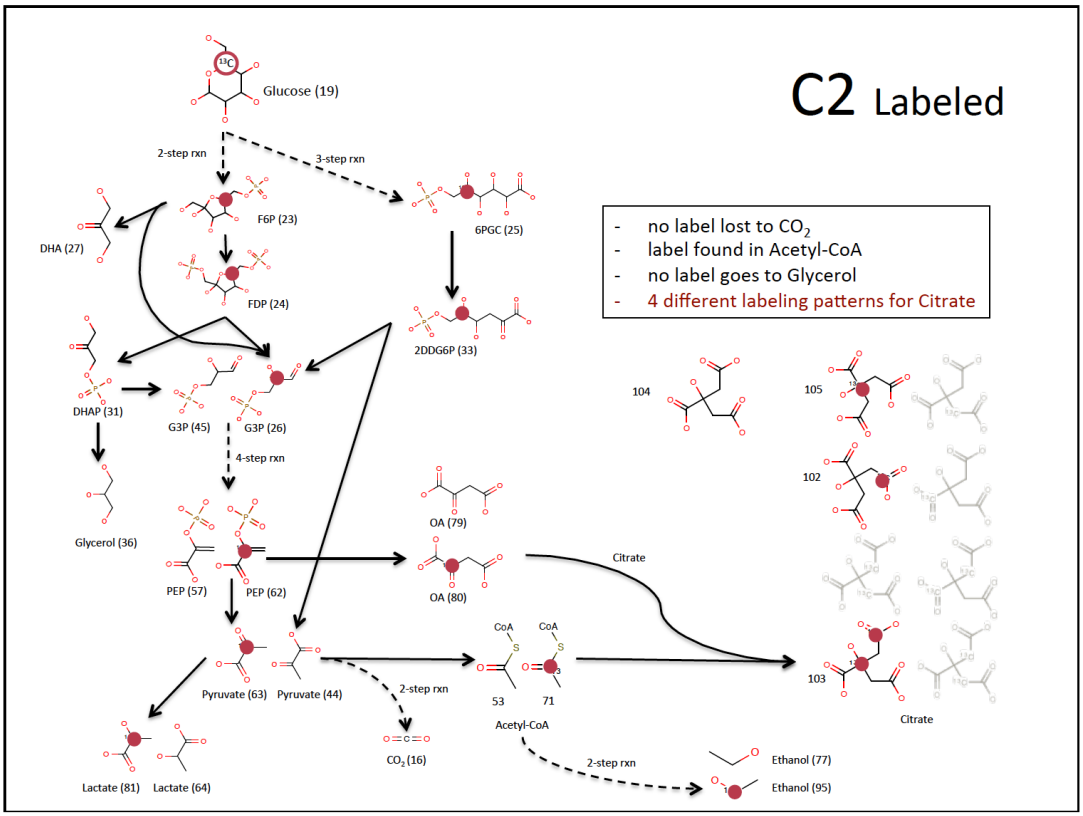
We continued our analysis by tracing the flow of labelled carbon atoms through the resulting atom-mapped networks. For a case study, we focused on the citrate molecule and we compared different obtained labelled patterns (isotopomers) for this compound depending on the position of labelled carbon atom in glucose. For six labelled glucose molecules, C1-C6, there were eight differently labelled and one unlabelled structure of

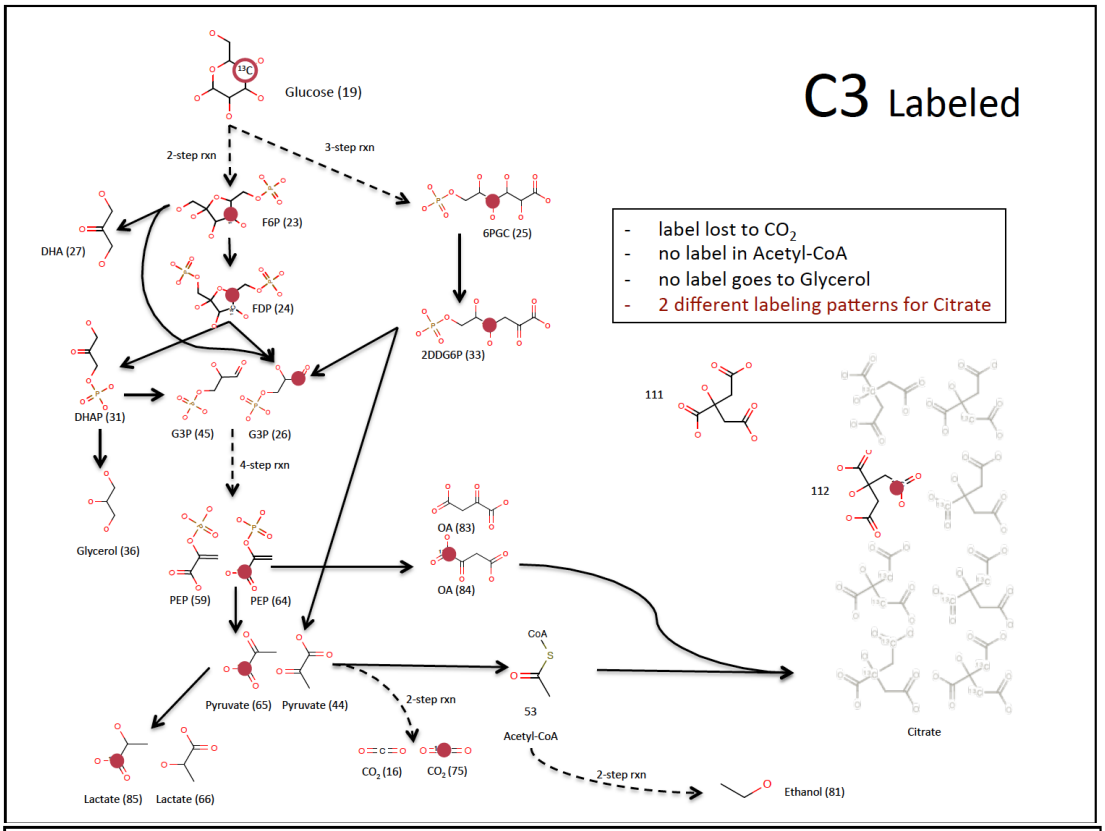
citrate, i.e. in total, there were nine out of 16 possible isotopomers for this compound. We visualized the flow of the labelled carbon towards citrate for six labelling patterns of glucose C1 to C6 in Figure 6.6.

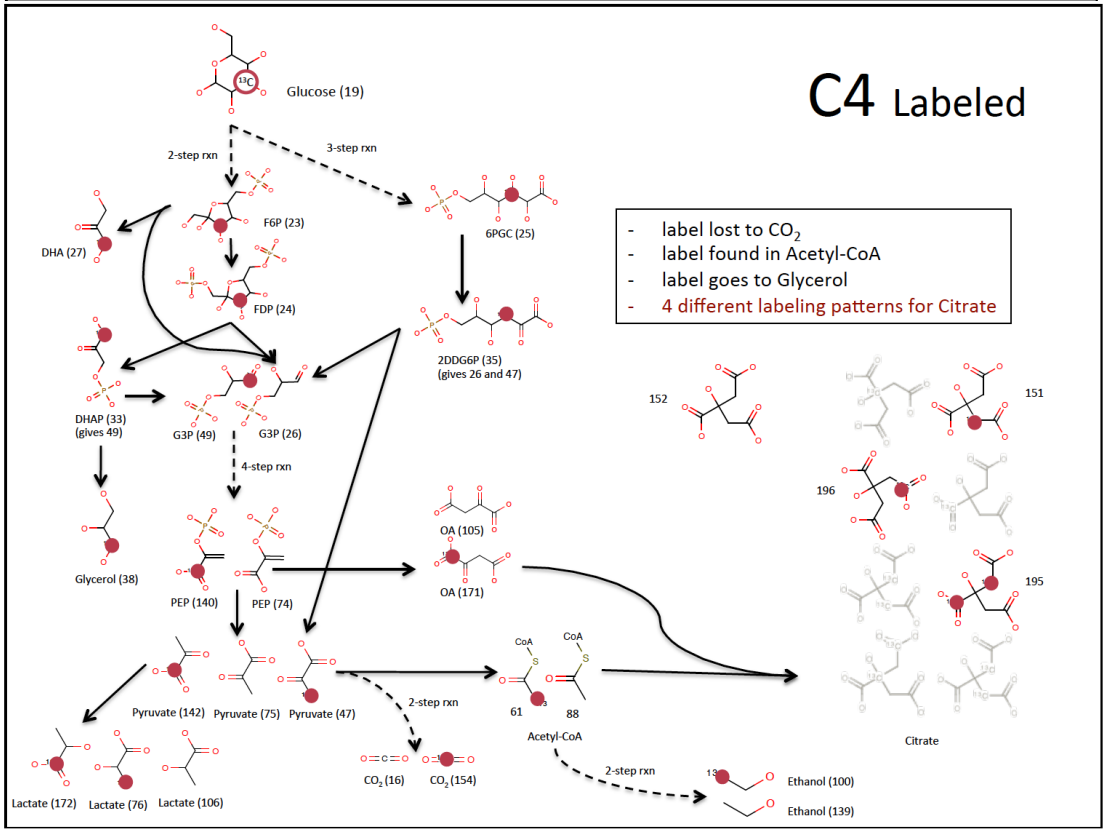
The visualisation of atom-mapped networks (Figure 6.6) allowed us to assess how different parts of metabolic networks such as Glycolysis, TCA cycle, etc. contributed to the diversity of generated isotopomers. For instance, while the pentose phosphate pathway did not largely contribute to the scrambling of carbon labels, glycolysis provided multiple alternative pathways for the carbon atoms, which resulted in a high number of isotopomers for the compounds at the end of glycolysis.

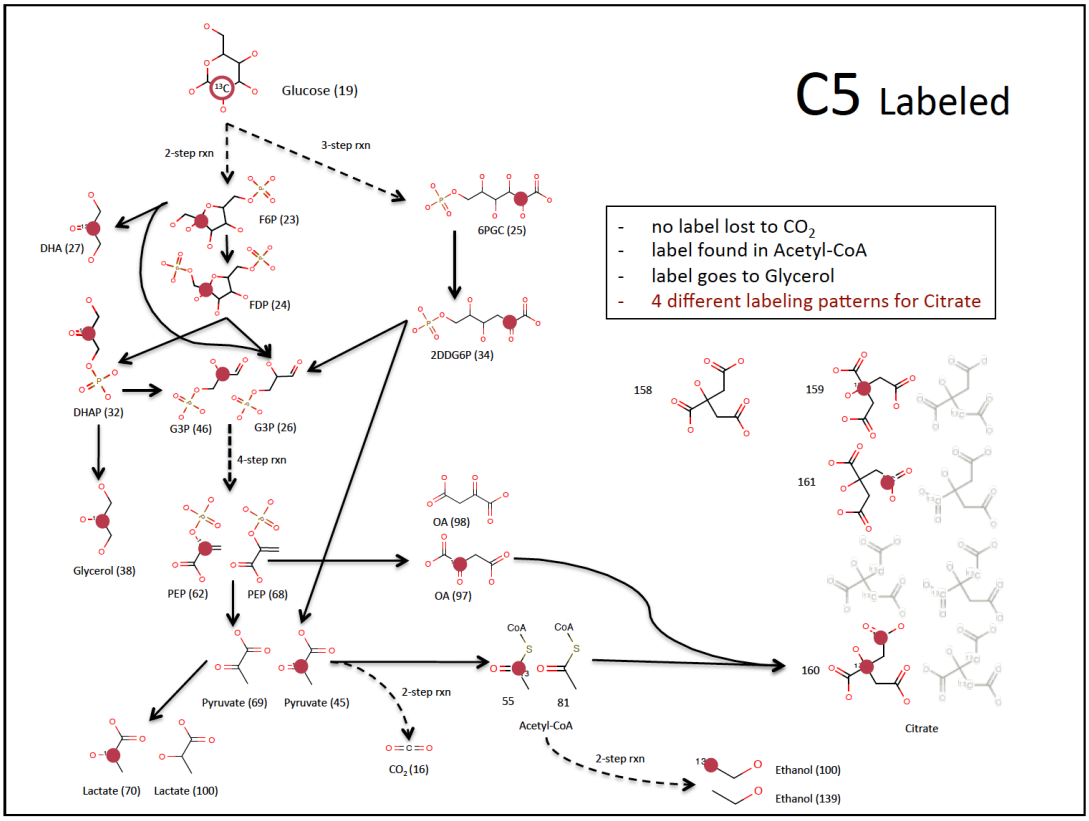
We further observed that some of the compounds like 2-oxoglutarate, glutamine, pyruvate and glutamate were highly connected despite a relatively small number of isotopomers. This was due to the fact that these compounds did not only react as substrates, but they were involved in some reactions as cofactors. In our visualisation, we excluded all cofactors that did not participate in carbon transfer (i.e. those that did not act as both cofactors and substrates) because of their high connectivity the network analysis became too involved.


C1 Labeled
Glucose (19)
2-step rxn
3-step rxn
F6P (23)
6PGC (25)
DHA (27)
FDP (24)
2DDG6P (33)
DHAP (31)
G3P (45)
G3P (26)
4-step rxn
Glycerol (36)
OA (84)
OA (85)
PEP (59)
PEP (64)
Pyruvate (65)
Pyruvate (44)
2-step rxn
CoA
CoA
53
75
Acetyl-CoA
CO_2 (16)
2-step rxn
Ethanol (82)
Ethanol (105)
Lactate (86)
Lactate (66)
115
113
114
Citrate
- no label lost to CO_2
- label found in Acetyl-CoA
- no label goes to Glycerol
- 3 different labeling patterns for Citrate


C2 Labeled
Glucose (19)
2-step rxn
3-step rxn
F6P (23)
6PGC (25)
DHA (27)
FDP (24)
2DDG6P (33)
DHAP (31)
G3P (45)
G3P (26)
4-step rxn
Glycerol (36)
OA (79)
OA (80)
PEP (57)
PEP (62)
Pyruvate (63)
Pyruvate (44)
2-step rxn
CoA
CoA
53
71
Acetyl-CoA
CO_2 (16)
2-step rxn
Ethanol (77)
Ethanol (95)
Lactate (81)
Lactate (64)
104
105
102
103
Citrate
Citrate
- no label lost to CO_2
- label found in Acetyl-CoA
- no label goes to Glycerol
- 4 different labeling patterns for Citrate


C3 Labeled
Glucose (19)
2-step rxn
3-step rxn
F6P (23)
6PGC (25)
DHA (27)
FDP (24)
2DDG6P (33)
DHAP (31)
G3P (45)
G3P (26)
4-step rxn
Glycerol (36)
OA (83)
OA (84)
PEP (59)
PEP (64)
Pyruvate (65)
Pyruvate (44)
2-step rxn
CoA
53
Acetyl-CoA
Citrate
Lactate (85)
Lactate (66)
CO2 (16)
CO2 (75)
2-step rxn
Ethanol (81)
- label lost to CO2
- no label in Acetyl-CoA
- no label goes to Glycerol
- 2 different labeling patterns for Citrate
111
112


C4 Labeled
Glucose (19)
2-step rxn
3-step rxn
F6P (23)
6PGC (25)
DHA (27)
FDP (24)
2DDG6P (35)
(gives 26 and 47)
DHAP (33)
(gives 49)
G3P (49)
G3P (26)
4-step rxn
Glycerol (38)
OA (105)
OA (171)
PEP (140)
PEP (74)
Pyruvate (142)
Pyruvate (75)
Pyruvate (47)
2-step rxn
CoA
61
88
Acetyl-CoA
Citrate
Lactate (172)
Lactate (76)
Lactate (106)
CO2 (16)
CO2 (154)
2-step rxn
Ethanol (100)
Ethanol (139)
- label lost to CO2
- label found in Acetyl-CoA
- label goes to Glycerol
- 4 different labeling patterns for Citrate
152
151
196
195


C5 Labeled
Glucose (19)
2-step rxn
3-step rxn
F6P (23)
6PGC (25)
DHA (27)
FDP (24)
2DDG6P (34)
DHAP (32)
G3P (46)
G3P (26)
4-step rxn
OA (98)
Glycerol (38)
PEP (62)
PEP (68)
OA (97)
Pyruvate (69)
Pyruvate (45)
2-step rxn
CoA
CoA
55
81
Acetyl-CoA
Lactate (70)
Lactate (100)
CO2 (16)
2-step rxn
Ethanol (100)
Ethanol (139)
- no label lost to CO2
- label found in Acetyl-CoA
- label goes to Glycerol
- 4 different labeling patterns for Citrate
158
159
161
160
Citrate

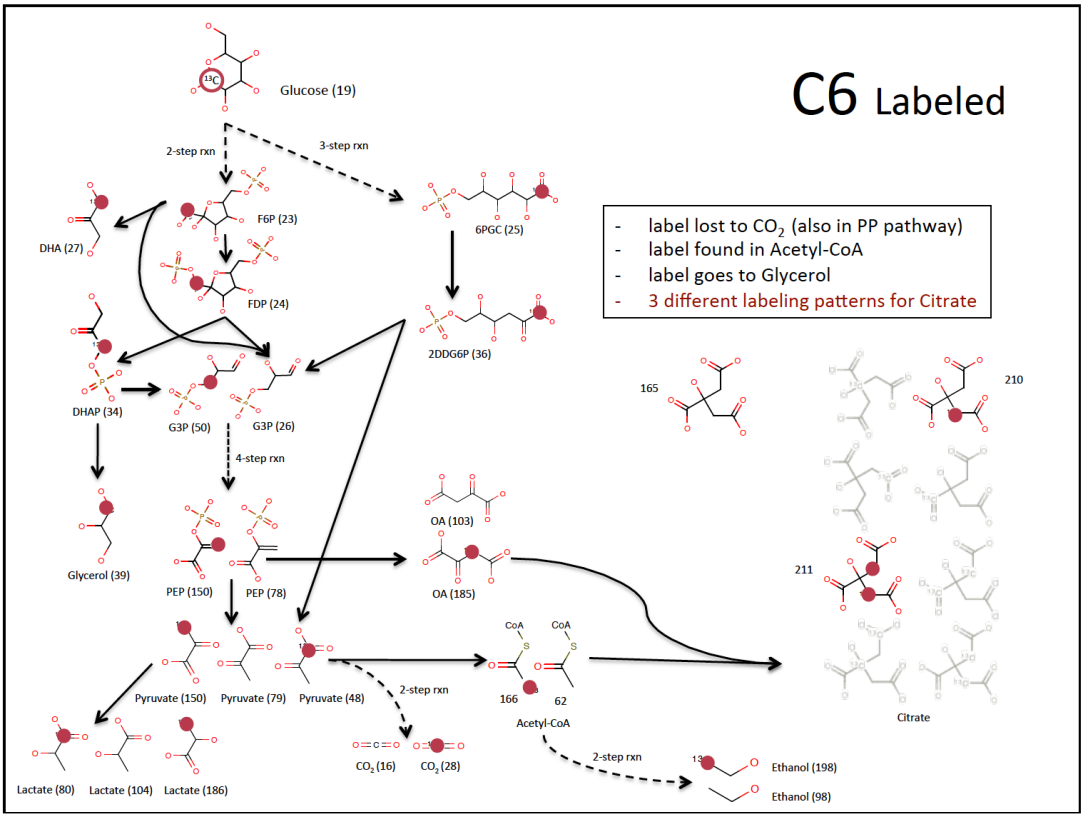

C6 Labeled
Glucose (19)
2-step rxn
3-step rxn
F6P (23)
6PGC (25)
DHA (27)
FDP (24)
2DDG6P (36)
DHAP (34)
G3P (50)
G3P (26)
4-step rxn
OA (103)
Glycerol (39)
PEP (150)
PEP (78)
OA (185)
Pyruvate (150)
Pyruvate (79)
Pyruvate (48)
2-step rxn
CoA
CoA
166
62
Acetyl-CoA
Lactate (80)
Lactate (104)
Lactate (186)
CO2 (16)
CO2 (28)
2-step rxn
Ethanol (198)
Ethanol (98)
- label lost to CO2 (also in PP pathway)
- label found in Acetyl-CoA
- label goes to Glycerol
- 3 different labeling patterns for Citrate
165
210
211
Citrate

Figure 6-6: Section of the atom-mapped *E. coli* core metabolism is presented for single *in silico* labelled carbon of glucose, changing from C1 to C6. Red dots indicate the positions of ^{13}C arising from the initial labelled glucose and conserved in the pathway. Varying the position of the labelled carbon atom on glucose results in different citrate isotopomers, which are highlighted for each case. Dotted lines represent the combination of several reactions steps that do not affect the position of labelled carbon. The number on the bottom of each compound corresponds to its entry in the generated network (graph node).

6.4 Conclusions

In this study, we presented a novel, systematic approach capable of reproducing the known metabolic knowledge with an extra level of information - the atom-mapped data. We performed the atom mapping of biological reactions, pathways and networks using enzymatic reaction rules that reflect a biochemical reality, therefore chemical and biological validity of the reported atom-mapped reactions are guaranteed.

We demonstrated our results through the reconstruction of an atom-level representation of the *E. coli* core metabolism and we elucidated the flow of single carbon atoms through the network starting from glucose as the carbon source. Furthermore, we analysed the generated atom-mapped networks from different aspects such as the growth in number of isotopomers or the loss of the isotopic tracer to CO_2. This information is very useful for optimizing the design of ^{13}C-MFA experiments. The result of this study shows precisely which part of metabolism is contributing more for the production of biomass building blocks.

Our focus in this study was on the analysis of *carbon* flow through the metabolic pathways, nevertheless similar studies exploring the metabolic fate of atoms other than carbon could reveal new insights into the internal organization of metabolism. The proposed method allows also reconstructing the atom-mapped metabolic networks in an extended *search space*. This way, for example, we could consider all KEGG reactions for a reconstructed *E. coli* network and study how the integration of reactions non-native to *E.coli*, but which exist in other organisms, would affect the carbon flow in the network. In addition, integrating novel pathways generated by the original BNICE.ch framework in the search space could further increase the potential of this type of studies. Such results may subsequently be used: (i) for applications in synthetic biology;

(ii) for generating hypotheses regarding alternative or missing pathways in metabolism; and (iii) to make detailed predictions for the simulation of MFA experiments.

The proposed method is capable of handling big metabolic networks containing hundreds of compounds and reactions. This feature distinguishes the proposed atom mapping method from other available algorithms as it opens up a possibility for the reconstruction of genome-scale atom-mapped metabolic networks.

Chapter 7

THERMODYNAMICS of METABOLISM

How energetics constraint biochemistry?

SUMMARY

In this chapter we present a coherent and unified developed framework for the estimation of thermodynamics of metabolism and the rationalization of feasibility analysis of metabolic pathways. This work is the first phase of an on going project on thermodynamics of metabolism using GCM methods coupled with the quantum thermochemical calculations.

The lack of the available *ab initio* methods for the direct calculation of thermodynamic properties at aqueous solutions (which correspond to biological systems) limits their applications in biological studies. To address this limitation, our proposed pipeline first calculates the thermodynamic properties in the gas phase and the results of this study coupled with the corresponding hydration/solvation results (on going project based on this chapter) would result in the desired thermochemical quantities in aqueous solutions.

In this chapter we introduce our methods and present our results for the estimation of thermodynamic properties for a wide range of metabolites. Furthermore, we compare our calculations with reliable experimental measurements and predictive calculations from the literature, when available.

7.1 Introduction

Intrinsic properties of metabolites from gas-phase quantum chemical calculations form the basis for understanding their behaviour in more complex systems such as proteins, nucleic acids and polysaccharides. Remarkable progress is made for small to moderate-sized compounds with up to ten non-hydrogen atoms. However, the biochemical compounds in metabolic pathways such as aminoacids and oligopeptides, saccharides, nucleosides and their derivatives, often exceed this molecular size limit. For the heavier metabolites, we further need computational information since reliable experimental information for them is rather limited and often non-existing. Most of the compounds of interest in this work are solid / crystalline at ambient conditions with high melting points and very low vapour pressures and, thus, are elusive to direct gas-phase measurements.

Most often, the standard heat of formation of these compounds in the gaseous state are obtained by combining the heats of formation in the pure solid state (typically, from heats of combustion) and the heat of sublimation (typically, by extrapolation from values obtained at significantly higher temperatures – if the compounds are not decomposed). Both of these typical experimental measurements, however, require extreme care as they are prone to significant experimental errors [197]. Therefore, for many of the heavy metabolites described above, the basic thermochemical quantities in the ideal gas state still resist an experimental determination with the acceptable "chemical accuracy" of 1 kcal/mol. Attempts to develop predictive group contribution methods based on experimental results for smaller molecules [198,199] are not always successful for the above types of metabolites and require significant improvements [200]. Thus, today there is much interest in accurate theoretical calculations of thermodynamic quantities of these metabolites.

Moreover, most of the thermodynamic studies in the literature are confined to ambient conditions. Their extension to remote temperatures and pressures requires accurate knowledge of additional properties relevant to the interaction of the studied metabolites with their neighbouring molecules. Heat capacities, thermal expansivities and isothermal compressibilities are among the properties that must be known for such an extension. In addition, the effect of external conditions on solvation phenomena,

primarily hydration phenomena, is of key importance in the evolution of biochemical reactions and processes at remote conditions. This includes metabolite dissociation, ligand-binding, protein folding, and ion distributions. In this regard, molecular thermodynamics, which combines quantum chemical calculations with classical mixture thermodynamics and equation-of-state approaches for extrapolation to high temperatures and pressures, can be used for understanding the metabolism of microorganisms, not only at ambient conditions but also at extreme conditions of temperature and pressure [201-210].

In this chapter, we introduce our cohesive approach for the reliable estimation of the basic thermochemical quantities of metabolites, biomolecules and associated biochemical reactions over an extended range of external conditions of temperature and pressure. As for the results, we focus on quantum chemical calculations of enthalpies and free energies of formation in gas phase via *isodesmic reactions* of key metabolites, such as aminoacids/oligopeptides, oligosaccharides, and nucleotides /nucleosides, for which experimental information is difficult to obtain. The obtained quantum chemical calculations complement the available thermochemical compilations and form the basis for the expansion and testing of subsequent developments.

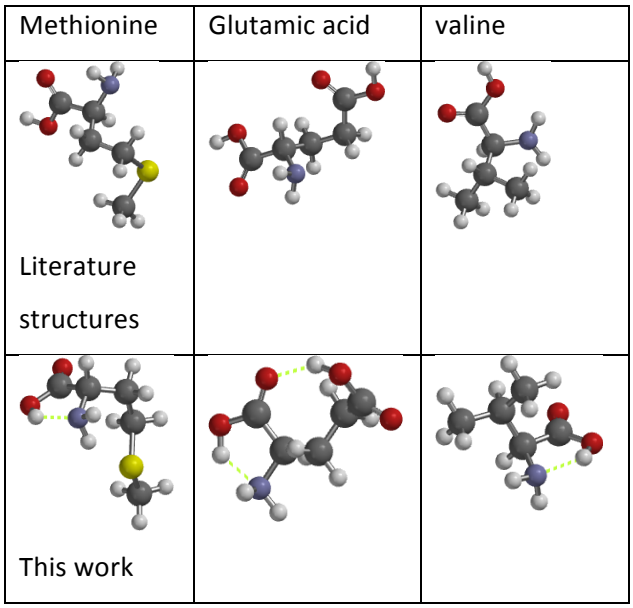
7.2 Methods

7.2.1 Accounting for conformers, anomers and tautomers

In figure 7.1, we compared the most stable structures obtained for methionine, glutamic acid and valine with the corresponding structures proposed in [157], and the most stable structure for open chain glucose with the corresponding structures proposed in [211]. The differences in the calculations are not significant in the case of aminoacids. What is worth pointing out is the different stabilizing factors in the two cases. As is clear from the figure, the second row structures (this work) exhibit intramolecular hydrogen bonding that includes the relatively strong OH--- NH bond. In contrast, the upper row structures do not exhibit such intramolecular hydrogen bonds and appear as less compact structures. Based on these properties we expect that the upper row structures prevail at higher temperatures while the lower row structures be dominant at lower temperatures. In the case of glucose chains, we observe that both structures are

stabilized by four intramolecular hydrogen bonds of the same OH--- O type and the difference in the calculations is not negligible. The right structure (this work) is more stable by ca. 9 kJ/mol. The reason for this difference resides on the more cooperative character of hydrogen bonding in the right structure (four vs. three consecutive hydrogen bonds)[212] and the relatively open structure (free terminal OH).

A



B

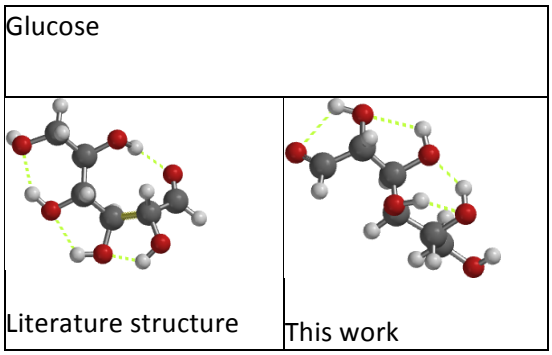


Figure 7-1: **A**: Comparison of most stable structures for methionine, glutamic acid and valine with the corresponding structures proposed by Stover et al.[157]. **B:** Comparison of the most stable structure of glucose chain with the corresponding literature structure[211].

From the thermodynamic point of view, the pertaining value for the metabolite is an *average* over the conformer population considered coexisting at equilibrium. This averaging is done by adopting the classical Boltzmann distribution equation [213]:

$$X_i = \frac{exp\left(-\frac{\Delta_f G_i}{RT}\right)}{\sum_j exp\left(-\frac{\Delta_f G_j}{RT}\right)} \qquad (7.1)$$

where X_i is the mole fraction of conformer i at equilibrium and $\Delta_f G_i$ is the Gibbs free energy of formation of conformer i. The summation is over all conformers of the studied molecule. The contribution of each conformer to the overall heat of formation of the studied metabolite is just the product of conformer's heat of formation with its mole

fraction at equilibrium, and similarly for the Gibbs free energy of formation. Using equation (8.1), we observe that this contribution falls rapidly as the conformer's Gibbs free energy of formation departs from its lowest value in the conformer set, or, as the relative Gibbs free energy of formation departs from zero (Figure 7.2). Note that the relative Gibbs free energy corresponds to the difference of Gibbs free energy of formation of conformer i from the corresponding quantity of the global minimum or most stable conformer, i.e.

$$\Delta\Delta_f G_i = \Delta_f G_i - \Delta_f G_{min} \tag{7.2}$$

Thus, although the heat of formation of some conformers may depart by as much as 30 kJ/mol from the corresponding lowest value, the so-calculated average heat of formation differs by less than 4.0 kJ/mol from the lowest conformer's value.

The above conformer distribution analysis is valid as long as the conformer population is known sufficiently well. Omitting conformers near the global minimum may lead to significant errors. However, omitting all conformers above the usual cut-off of 1 kcal/mol in relative Gibbs free energy may also introduce a non-negligible error since the omitted conformers may be numerous. Another source of error is the fact that Gibbs free energies of formation are not usually known with high precision and the above averaging is then done with the overall energy.

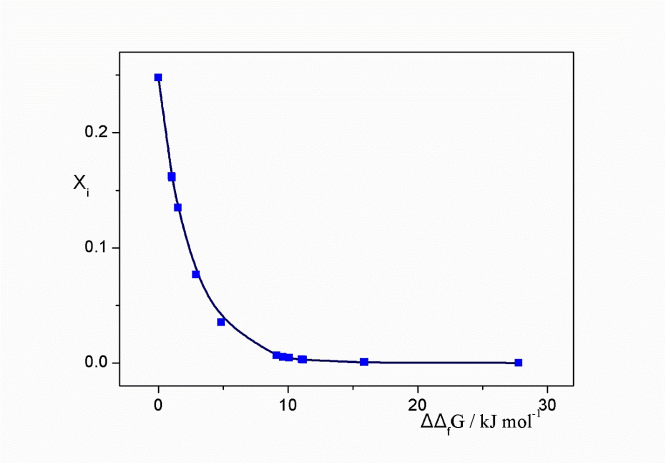


Figure 7-2: The mole fraction of conformers at equilibrium as a function of the corresponding relative Gibbs free energy of formation for aspartic acid.

In the case of conformers with extensive intramolecular hydrogen bonding, which is often the case with heavy metabolites, this error may not be negligible either.

The situation may be different in the case of *tautomers*. As figure 7.3 shows, for the calculation of the thermochemical quantities for inosine, as an example, we need the corresponding quantities for hypoxanthine. Hypoxanthine may be found in 7 tautomeric forms including the enol form of 6-hydroxypurine.

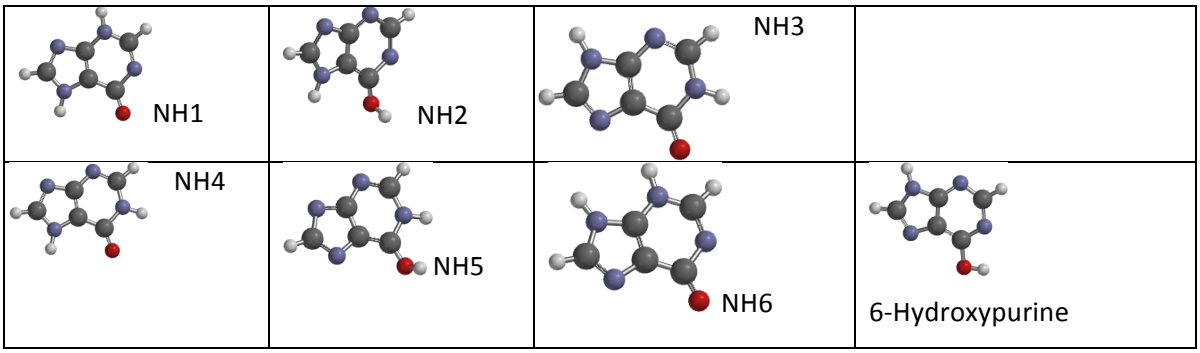


Figure 7-3: The tautomers of hypoxanthine

These structures correspond to a well-known isomer population for which equation (1) may be applied unequivocally. One may argue that some of these tautomers are not relevant since the ribose group occupies one site in inosine. However, regarding hypoxanthine itself, its average thermochemical quantities may still be obtained by applying equation (1), leading to a value of just 0.4 kJ/mol higher than the heat of formation of the most stable conformer, NH4. The situation in tautomers of nucleosides or histidine is different since some of them may be favored (stabilized) by strong intramolecular hydrogen bonds. The averaging results for guanine and cytosine are similar. Although one may infer from this discussion that the thermochemical quantities of the most stable tautomer are good approximations for the corresponding quantities for the metabolite, the knowledge of these quantities for the less stable tautomers are also important.

The case of sugar *anomers* with open and cyclic structures is again different. The results through detailed sugar puckering analyses[211,214] show that particular puckering geometries may be favoured in various glycobiology processes over the usual low

energy equatorial $4C_1$ conformation and, thus, a detailed knowledge of the puckering landscape is crucial for understanding these processes.

Thus, it makes more sense to focus on the thermochemical quantities of the global minimum in each class of structures (open chain, furanoses, pyranoses) and use the above puckering analyses for the averaging via equation (8.1). However, considering the reported range and the distribution of enthalpy differences in the above puckering analyses, it is clear that the thermochemical quantities of sugars do not deviate significantly from the corresponding quantities of the global minimum in each case.

7.2.2 Selection of isodesmic reactions for calculation of heats and Gibbs free-energies of formation

Apart from the standard requirement for the preservation of number and type of bonds on both sides of the isodesmic reaction, heavy metabolites may pose additional requirements as they may exhibit extensive intramolecular hydrogen bonding. One may wonder whether the reactants should account for the intramolecular interactions of the metabolites.

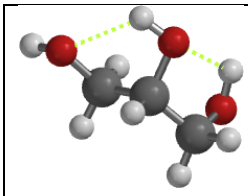
Glycerol is a good example for the case. Glycerol may exhibit two intramolecular hydrogen bonds, which further stabilize its conformers. Simple conformer search shows that conformer #1 in figure 7.4 is the most stable glycerol conformer. Two alternative isodesmic reactions for obtaining glycerol are the followings:

2 ethanol + methanol => glycerol + 2 methane (A)

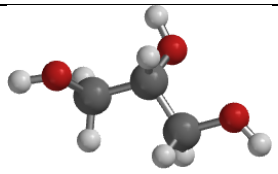
or

2 ethylene glycol => glycerol + methanol (B)

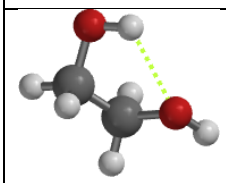
In reaction B, we also have the option to choose between alternative reactant (ethylene glycol) conformers, as shown in figure 5.



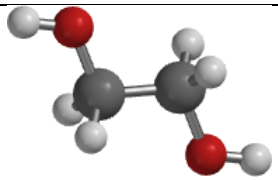
Glycerol – conformer #1



Glycerol – conformer #2



Ethylene glycol – conformer #1



Ethylene glycol - conformer #2

Figure 7-4: Conformers of glycerol and ethylene glycol with and without intramolecular hydrogen bonds (shown by dashed lines).

Table 7.1 summarizes the predicted thermochemical quantities with calculations at the DFT-D3/B3LYP/def2-TZVPD level for the various alternative reactions or reactant conformer combination schemes. As shown in this table, the isodesmic reaction A could lead to acceptable predictions. However, the reaction heats (and free energies) are not negligible. This may entail some error from the inefficient compensation of theory flaws from both reaction sides. As expected, the use of isodesmic reaction B with ethylene glycol conformer #2 leads to unacceptable results. In contrast, the intramolecularly hydrogen-bonded conformer #1 of ethylene glycol is the most appropriate reactant for isodesmic reaction B as shown by the predicted results in table 1, which also conforms to experiment. The heat of reaction in this case (-4.9 kJ/mol) is small and is probably indicative of a further stabilization of glycerol conformer #1 due to cooperativity of its dual intramolecular hydrogen bonding.

The glycerol example will be used as a guide for the selection of isodesmic reactions and conformer reactants in this work and especially in the case of saccharides.

The isodesmic reactions that have been considered in the present work are shown in Table 7.2.

Table 7-1: The predicted thermochemical quantities for glycerol based on quantum chemical calculations at the DFT-D3/B3LYP/def2-TZVPD level.

Table 7-2: The isodesmic reactions for neutral metabolites

Metabolite	Isodesmic Reaction
Aminoacids / dipeptides	
H2NCHRCOOH	H2NCHRCOOH (Gly) + CH3R => H2NCHRCOOH + CH4
Arginine (Arg)	Gly+guanidine+butylamine => Arg + CH4 + NH3
Histidine (His)	Gly+imidazole+propane => His+2 CH4
Proline (Pro)	pyrrolidine+acetic acid => Pro+ CH4
Tryptophan (Try)	Gly+indol+propane => Try+2 CH4
Alanylalanine (AlaAla)	2Ala +CH3NH2 => AlaAla + CH3OH + NH3

Glycerol conformer	**Heat of reaction (kJ/mol)**	**Heat of formation (kJ/mol)**		**Free energy of formation (kJ/mol)**	
		Predicted	**Experim.**	**Predicted**	**Experim.**
Isodesmic reaction A					
#1	-62.0	-583.7	-577.9[1], -582.8[2]	-445.8	-448[2]
#2	-48.5	-570.0		-435.6	
Isodesmic reaction B, ethylene glycol conformer #1					
#1	-4.9	-582.6	-577.9[1], -582.8[2]	-446.0	-448[2]
#2	+8.7	-569.1		-435.7	
Isodesmic reaction B, ethylene glycol conformer #2					
#1	-22.9	-617.7	-577.9[1], -582.8[2]	-462.3	-448[2]
#2	-9.3	-604.1		-452.0	

Alanylglycine (AlaGly)	Ala + Gly + CH3NH2 => AlaGly + CH3OH + NH3
Alanylphenylalanine (AlaPhe)	Ala + Phe + CH3NH2 => AlaPhe + CH3OH + NH3
Glycylglycine(GlyGly)	2Gly + CH3NH2 => GlyGly + CH3OH + NH3

Phenylalanylglycine (PheGly)	Gly + Phe + CH3NH2 => PheGly + CH3OH + NH3
Glycylvaline (GlyVal)	Gly + Val + CH3NH2 => GlyVal + CH3OH + NH3
Leucylglycine (LeuGly)	Leu + Gly + CH3NH2 => LeuGly + CH3OH + NH3
Saccharides	
αD-Glucose (αDGl)	3HOCH2CH2OH+tetrahydropyran(THP) => αDGl + CH3OH + 2C2H6
βD-Glucose (βDGl)	3HOCH2CH2OH + THP => βDGl + CH3OH + 2C2H6
Glucose – chain (GlCh)	3HOCH2CH2OH + C2H6 + CH3CHO => GlCh + CH3OH + 3CH4
2-deoxy-D-Glucose (2-deGl)	2HOCH2CH2OH + THP => 2-deGl + CH4 + C2H6
βD –Galactose (βDGal)	3HOCH2CH2OH + THP => βDGal + CH3OH + 2C2H6
βD –Mannose (βDMan)	3HOCH2CH2OH + THP => βDMan + CH3OH + 2C2H6
βD –Ribose (βDRib)	2HOCH2CH2OH + THF => βDRib + CH4 + C2H6
αD-Ribofuranose (αDRib)	2HOCH2CH2OH + THF => αDRib + CH4 + C2H6
2–deoxyD-ribose(2-deRib)	2HOCH2CH2OH + THF => 2-deRib + CH3OH + C2H6
Fructose-chain (FrCh)	3HOCH2CH2OH + CH3COCH3 => FrCh + CH3OH + 2CH4
βD-Fructopyranose (βDFrp)	3HOCH2CH2OH + THP => βDFrp + CH3OH + 2C2H6
βD-Fructofuranose (βDFr)	3HOCH2CH2OH + THF => βDFr + CH3CH2OH + 2CH4
βD–Xylofuranose (βDXyf)	3HOCH2CH2OH + THF => βDXyf + CH3CH2OH + CH3OH + C2H6
βD–Xylopyranose (βDXyp)	2HOCH2CH2OH + THP => βDXyp + 2 C2H6
αD–Xylopyranose (αDXyp)	2HOCH2CH2OH + THP => αDXyp + 2 C2H6
Sucrose	βDGl + βDFr + CH3COCH3 => Sucrose + 2 CH3OH
Maltose	2βDGl + CH3COCH3 => Maltose + 2 CH3OH
Nucleobases/Nucleosides	
6-Hydroxypurine	Adenine + CH3OH => 6-Hydroxypurine + CH3NH2
Adenosine	Adenine + βDRib + CH3NH2 => Adenosine + CH3OH + NH3
2-Deoxyadenosine	Adenine + 2-deRib + CH3NH2 => 2-Deoxyadenosine + CH3OH + NH3
2,3-Dideoxyadenosine	Adenosine + 2 CH4 => 2,3-Dideoxyadenosine + 2 CH3OH
Inosine	Hypoxanthine + βDRib + CH3NH2 => Inosine + CH3OH + NH3
Guanosine	Guanine + βDRib + CH3NH2 => Guanosine + CH3OH + NH3
Cytidine	Cytosine + βDRib + CH3NH2 => Cytidine + CH3OH + NH3
Thymidine	Thymine + βDRib + CH3NH2+ CH4 => Thymidine + 2CH3OH + NH3

Deoxythymidine	Thymine + THF+ CH3NH2+ CH3CH2OH => Deoxythymidine + 2CH4 + NH3
Uridine	Uracil+ βDRib + CH3NH2 => Uridine + CH3OH + NH3

7.2.3 Thermochemical calculations for neutral metabolites

We calculated the heats and Gibbs free energies of the isodesmic reactions (Table 7.2) at 298.15 K and ambient pressure. We added ZPE (zero point vibrational energy) to SPE together with the thermal corrections for the enthalpy or the Gibbs free energy for each reactant or product. This results in their total enthalpy, H^0, and Gibbs free energy, G^0:

$$H^0 = SPE + \alpha ZPE + H_v \quad (7.3)$$

and

$$G^0 = SPE + \alpha ZPE + G_v \quad (7.4)$$

The scaling factor, α, has values close to 1. In this work it was set equal to 0.975. The heat of a reaction:

$$\nu_1 R_1 + \nu_2 R_2 \Rightarrow \nu_3 R_3 + \nu_4 R_4$$

can then be obtained in the usual manner as

$$\Delta_r H^0 = \nu_3 H^0_3 + \nu_4 H^0_4 - \nu_1 H^0_1 + \nu_2 H^0_2 \quad (7.5)$$

A similar approach can be used for the free energy of the same reaction, $\Delta_r G^0$. We can also derive the same result from the heats of formations of reactants and products, or:

$$\Delta_r H^0 = \nu_3 \Delta_f H^0_3 + \nu_4 \Delta_f H^0_4 - \nu_1 \Delta_f H^0_1 + \nu_2 \Delta_f H^0_2 \quad (7.6)$$

and similarly for $\Delta_r G^0$.

Calculation of the standard heats and free energies of formation of a metabolite requires accurate knowledge of the corresponding properties for all reactants and products of the isodesmic reaction. By replacing the known quantities we can use equation (7.6) in order to obtain the unknown heat of formation (similarly the unknown Gibbs free energy of formation) of the studied metabolite. The existing compilations[147-149,215] provide reliable data for these computations, which are summarized in Appendix Table A3 for the compounds that participate in our isodesmic reactions.

7.2.4 Thermochemical calculations for ionic metabolites

Gas phase acidities, gas phase basicities (GB) and proton affinities (PA) are important elements for understanding the acid-base behavior of heavy metabolites not only in the

gas phase but also in aqueous solutions and numerous biological processes [216,217]. The extensive experimental and theoretical research over several decades has led to reliable compilations of the above properties, especially for aminoacids [150,152,153,157,158]. For this case, one can use these experimental data to estimate heats and free energies of formation of protonated and deprotonated aminoacids if the corresponding quantities for the neutral species are available. On the other hand, one can perform the same level of quantum thermochemical calculations as described in section 8.2.1, and use appropriate isodesmic reactions in order to estimate the thermochemical quantities of the ionic metabolites. We have followed both approaches and we compare the results in the next section. In order to clarify the computations, we first recall the definitions of gas phase acidity, GB, and proton affinity, PA.

Let reaction C be the protonation reaction of metabolite M :

$$M\,(g) + H^+\,(g) \Rightarrow MH^+\,(g) \qquad (C)$$

and reaction D the deprotonation reaction for the same metabolite:

$$M\,(g) \Rightarrow [M\text{-}H]^-\,(g) + H^+\,(g) \qquad (D)$$

The gas-phase basicity of metabolite M is equal to the negative of the standard Gibbs free energy change of reaction C, i.e.,

$$GB\,(M) = -\Delta_r G^0_{C'} \qquad (7.7)$$

and its proton affinity is equal to the negative of the standard enthalpy change of the same reaction C, i.e.,

$$PA\,(M) = -\Delta_r H^0_C \qquad (7.8)$$

The standard Gibbs free energy change of the deprotonation reaction D is the gas-phase acidity of metabolite M, i.e.,

$$\Delta G_{acid}\,(M) = \Delta_r G^0_D \qquad (7.9)$$

and similarly for the enthalpy:

$$\Delta H_{acid}\,(M) = \Delta_r H^0_D \qquad (7.10)$$

Calculations of gas phase acidities of aminoacids are typically done via the isodesmic reaction E

$$Aminoacid + CH3COO^- \Rightarrow [Aminoacid - H]^- + CH3COOH \qquad (E)$$

and by anchoring the calculations to the known gas phase acidity for acetic acid [152,156,159]. Alternatively, one may use benzoic acid in the isodesmic reaction E or even glycine as there exist consensus values for it [152], namely ΔG_{acid} (Gly) = 1403.5 ± 1.5 kJ/mol and ΔH_{acid} (Gly) = 1433.5 ± 1.5 kJ/mol.

Our calculations for ΔG_{acid} and ΔH_{acid} for glycine (1404.1 and 1434.4, respectively) fall within the consensus range and were used as a basis for the calculation of heats and free energies of formation of our deprotonated metabolites. For this purpose, however, we also need values for the heat and free energy of formation of deprotonated glycine itself. These quantities require the knowledge of corresponding values for proton (H^+) or any other cation. The adopted standard thermochemical values for proton in this work, $\Delta_f G^0(g, H^+)$ = 1513 kJ/mol and $\Delta_f H^0(g, H^+)$ = 1533 kJ/mol, were recommended by Truhlar et al [218] and were obtained by using the electron convention and the Fermi-Dirac statistics (not the Boltzmann statistics). Thus, the heat of formation of deprotonated glycine (cf. reaction D) is:

$$\Delta_f H \text{ (glycine-, g)} = \Delta H_{acid} \text{ (Gly)} - \Delta_f H^0(g, H+) + \Delta_f H \text{ (glycine, g)} = -491 \text{ kJ/mol}$$

and, similarly, $\Delta_f G$ (glycine-, g) = -409 kJ/mol. The thermochemical quantities of all other aminoacids can be obtained either directly as with glycine or through the isodesmic reaction

$$\text{Aminoacid} + \text{glycine-} \Rightarrow [\text{Aminoacid} - H]^- + \text{glycine} \qquad (F)$$

We further obtain the thermochemical quantities for the protonated metabolites in the same way.

7.3 Results and discussion

7.3.1 Thermochemical properties of neutral metabolites in ideal gas phase

We calculated the gas-phase standard heats and Gibbs free energies of formation of neutral metabolites and compared our estimations with available literature data, both, experimental and computational (Table 7.3). In general, our calculations for aminoacids fall within the range of literature values. They are also in agreement with the recent calculations by Stover et al [157]. On average, the discrepancy between the two sets of

calculations via isodesmic reactions is less than 1 kcal/mol. Our calculations are on average more negative, most probably due to the fact that our reference value for glycine (-391.7 kJ/mol), which seems to be more widely accepted [148,200,219,220], is lower than the value adopted by Stover et al [157] by almost 7 kJ/mol. Considering this different reference value, the two sets of calculations are well in accordance. However, our calculations still remain on average more negative than their G3-MP2 calculations. Stover et al. do not report Gibbs free energies of formation for comparison. Due to these discrepancies caused by the differences in the adopted reference values, we also report in Table 7.3 the heats and free energies of the isodesmic reactions that are not affected by such reference values.

Table 7-3: Estimated standard thermochemical properties (ideal gas phase) of neutral metabolites from isodesmic reactions (Table 7.2). Geometry optimization and frequency calculations at DFT-D3/B3LYP/def2-TZPV(D) level and energy (SPE) calculations at DFT-D3/B97-d/def2-QZPVD level. Values in parentheses were calculated at the same geometry but with SPEs at the DFT/ωB97X-d/6-311++G(2df,2p) level. All quantities are given in $kJmol^{-1}$.

Metabolite	$\Delta_r G$	$\Delta_f G$	$\Delta_r H$	$\Delta_f H$	
	estimated				**literature**
Aminoacids					
Alanine	-16.4	-298.0 (-298.0)	-21.0	-421.9 (-422.0)	-415.9[a] , -421.3[a] -414.7[g], -435.5[c], -415.9[h], -419.4[b]
Arginine	-42.0	-129.9	-63.4	-398.9	-389.1[a] ,-395.8[a], -380.6[b]
Asparagine	-10.8	-411.4 (-409.0)	-35.9	-612.0 (-610.0)	-591.8[a] , -610[a], -593.8[b] -609.1[c], -590.5[h]
Aspartic acid	-5.5	-621.8 (-622.7)	-19.6	-790.2 (-791.1)	-786.6[a] ,-793.3[a], -804.4[c], -786.7[h], -787.8[b]
Cysteine	-18.3	-272.8	-33.3	-396.5	-378.2[a], -395[a], -378.1[h], -382.6[n], - 397.1[n]
Glutamic acid	-21.4	-631.0 (-633.6)	-26.1	-819.0 (-821.7)	-807[a] , -815.9[a], -825.0[c], -807.4[h], -810.7[b]
Glutamine	-14.1	-395.8	-26.9	-626.0	-629.7[c], -621.7[a] -611.2[h], -618.1[b]
Glycine		-300.1[e]		-391.7	*-391.7*[e]. -392.1[f], -390.4[k]

Histidine	-56.9	-87.6	-75.9	-289.9	-221.8[h],-271.1[a] , -267.6[b] , -289.5[i]
Isoleucine	-16.0	-279.7 (-282.5)	-22.8	-493.5 (-496.3)	-486.6[a], -493.3[a], -487.1[b], -499.6[c], -486.8[h]
Leucine	-17.0	-280.6 (-282.7)	-24.5	-495.2 (-497.3)	-486.6 [a],-494.1[a], -489.7[b], -501.8[c], -486.8[f], -486.8[h], -487[5]
Lysine	-13.6	-205.9	-28.5	-458.8	-443.5[a] , -451.5[a], -443.4[h], -444.2[b]
Methionine	-16.2	-248.0 (-246.8)	-28.1	-427.8 (-426.4)	-412.1[a], -428.9[a], -426.5[b], -412.1[h]
Phenylalanine	-23.3	-142.2 (-139.0)	-36.3	-323.5 (-320.4)	-302.1[a] ,-322.6[a], -319.6[c], -312.9[f], -302[h], -318.4[b]
Proline	-12.4	-221.8	-29.8	-391.1	-373.2[a] , -387[a], -366.2[f,l], -373.3[h], -385.5[m]
Serine	-16.0	-433.5 (-433.6)	-29.5	-581.5 (-580.0)	-567.8 [a],-578.2[a], -567.8[h], -578.4[b]
Threonine	-16.5	-439.5 (-439.9)	-30.2	-620.0 (-620.3)	-603.8[a] ,-618.8[a], -620.9[c], -603.7[h], -616.0[b]
Tryptophan	-43.6	-29.8[*]	-58.0	-246.9	-217.2[a] ,-249.4[a], -215[h], -238.1[b]
Tyrosine	-17.2	-288.4 (-286.4)	-27.8	-489.0 (-487.0)	-482[a] , -490.4[a], -498.5[c], -481.9[h], -489.2[b]
Valine	-11.9	-282.2 (-283.4)	-21.3	-472.6 (-473.9)	-466.1[a] , -472.4[a], -481.2[c], -455.2[f], -466.1[h], -468.7[b], -455.2[j]
Dipeptides					
Alanylalanine	-15.1	-400.2	-18.4	-635.1	-648.3[c], -623.7[h]
Alanylglycine	-10.3	-397.6	-13.9	-600.4	-588.2[c]
Alanylphenylalanine	3.2	-226.1	4.8	-514.0	-534.9[b], -532.4[c] , -509.7[h]
Glycylglycine	-11.0	-400.4	-15.6	-571.8	-528[c], -571.9[h]
Phenylalanylglycine	-15.1	-246.5	-10.0	-498.0	-500.9[b]
Glycylvaline	-9.8	-381.3 (-378.4)	-15.0	-652.2 (-649.3)	-633.8[c], -647.9[h]
Leucylglycine	-15.5	-385.4	-22.6	-682.4	-652.2[c]
Nucleobases / Nucleosides					

Hypoxanthine	-21.8	172.8	-21.5	38.0	
6-Hydroxypurine	19.5	194.6	15.4	59.5	72.6[b]
Adenosine	-21.2	-123.0 (-128.7)	-31.6	-481.6 (-487.4)	-472.7[b]
2-Deoxyadenosine	29.3	18.2 (20.3)	40.1	-315.2 (-313.3)	-288.1[b]
2,3-Dideoxyadenosine	60.9	161.6	77.0	-151.9	-139.9[b]
Inosine	-12.0	-310.5 (-305.8)	-22.0	-659.7 (-655.1)	-646.4[b]
Guanosine	-12.1	-298.9	-22.5	-682.2	-664.2[b]
Uridine	-5.6	-675.3	-14.2	-991.5	-989.5[b]
Thymidine	30.2	-521.6 (-523.2)	30.6	-856.8 (-858.5)	
Deoxythymidine	-83.4	-373.8	-111.4	-691.5	-684.2[b]
Saccharides					
Cytidine	-12.7	-433.7	-18.9	-764.2	-768.9[b]
αD-Glucose	-74.5	-837.5	-91.2	-1115.0	
βD-Glucose	-74.6	-836.6	-89.1	-1112.8	-1040.1[e] -1113.5[b]
Glucose – chain	-78.0	-836.7	-109.1	-1102.5	-1016.2[e] -1095.5[b]
2-Deoxy-D-Glucose	-79.6	-682.7	-97.3	-941.9	-949.6[b]
βD –Galactose	-71.0	-833.0	-87.2	-1110.9	-1111.5[b]
βD –Mannose	-72.2	-834.2	-88.1	-1111.8	
βD-Fructopyranose	-89.1	-851.1 (-856.3)	-106.8	-1130.5 (-1136.9)	-1139.4[b]
βD-fructose (βD-fructofuranose)	-216.2	-934.8	-146.0	-1112.5	-1039.3[e]
Fructose-chain	-32.8	-828.6 (-829.0)	-60.5	-1093.2 (-1095.3)	-1113.3[b]
βD –Ribose	-79.6	-682.1	-100.1	-902.9	
αD-Ribofuranose	-78.5	-681.0	-102.2	-905.0	-907.9[b]
2-deoxy-D-Ribofuranose	-56.9	-547.5	-71.1	-747.6	
βD-Xylofuranose	-59.1	-684.5 (-572.4)	-76.6	-907.5 (-911.1)	-918.5[b]
βD-Xylopyranose	-72.0	-693.7	-84.8	-920.1	
αD-Xylopyranose	-72.6	-694.3	-86.6	-921.9	
Sucrose	45.9	-1419.2	33.2	-1985.2	
Maltose	3.9	-1459.2	-1.3	-2013.2	-1838.4[e]

[a]Stover 2012[157] ; [b]T1 (G3-MP2) (Spartan14, Wavefunction); [c]Domalski and Hearing 1993[199]; [d]Dorofeeva 2010[221]; [e]Goos database[148]; [f]Pedley 1994[222]; [g]Ngauv

Sabbah 1977[220]; [h]Sagadeev 2010[223], [i]Riffet [213], [j]NIST-JANAF [224], [k]Ngauv [220], [l]Sabbah [225]. [m]Contineanu[226]. [n]Ribeiro da Silva 2010[227].

Table 7.3 shows some inconsistencies between our calculations and the literature values. Direct comparison with the GCM predictions[199] cannot be made since they have used an erroneous value for glycine (-375.3 kJ/mol). The consistency with the composite T1 method (Spartan 14 database – Wavefunction) is very good in the case of PheGly but poor in the case of AlaPhe. However, our calculations for this latter dipeptide are in good agreement with the recent estimation by Sagadeev [200].

The literature on saccharides and nucleosides is not as rich as that of aminoacids. The focus on saccharides is on the more stable furanose or pyranose closed ring structures [211,214]. Their capacity to form multiple intramolecular hydrogen bonds, as shown in the previous section is contributing significantly to their stability and cannot be overlooked in selecting isodesmic reactions for their formation. There are very few literature data to compare with our calculations. Our calculations are consistent with the composite T1 predictions (Spartan14 dbase – Wavefunction) but we see noticeable differences with the reported ATcT values[148]. However, the discrepancies are too large (over 70 kJ/mol) to be attributed to erroneous calculations either via isodesmic reactions (our calculations) or via composite G3-MP2 method (Spartan 14). This is probably one of the very few cases to be reconsidered in ATcT compilation.

The nucleosides share features from nucleotides, especially tautomerism, and from the saccharide bD-ribose, especially the flexibility for intramolecular hydrogen bonding. As shown in figure 5, guanosine can form the strong OH –N hydrogen bonds with the guanine ring as well as the OH—O hydrogen bonds with ribose.

In uridine, the prevailing hydrogen bond is of the OH—O type with the carbonyl oxygen of the uracil ring. This hydrogen bonding capacity is a significant stabilizing factor for the nucleoside conformers and an adequate conformational search cannot overlook them.

For nucleosides, we made a thorough comparison in Table 4 with composite T1 calculations (Spartan 14 database – Wavefunction). The observed discrepancy between

Spartan 14 and our estimations for adenosine, guanosine, inosine and derivatives is, in part, stemming from the fact that for adenine base, Spartan 14 database gives a heat of formation of 276.3 kJ/mol, which is significantly higher than the ATcT database value of 225.7 kJ/mol, while for guanine it gives a value of 44.09 kJ/mol, which is again higher than the ATcT value of 16 kJ/mol. The inconsistencies for the other nucleobases are significantly smaller as shown in the calculations.

7.3.2 Thermochemical properties of ionized metabolites in ideal gas phase

In the majority of metabolites, the populations of species in the gas phase are dominated by the neutral as they are much more stable than the ionized structures. However, we also considered ionized structures, as this will contribute to our understanding of their hydration that will be important in the next phases of the project for bringing the gas phase estimation to aqueous phase, which is the case for biological systems. Our main focus on ionic structures is on aminoacids for which we considered both protonation and deprotonation reactions. In Table 7.4 we report the gas-phase acidities (reaction enthalpies and Gibbs free energies) for aminoacids as well as the standard enthalpies and Gibbs free energies of formation of the *deprotonated* aminoacids. The agreement of our calculations with literature data is satisfactory. We can however observe noticeable differences such as for the case of ΔG_{acid} for methionine (18 kJ/mol). The formation quantities are all negative but there are no literature data for comparison.

Table 7-4: Estimated gas-phase *acidities* and standard thermochemical properties of ionic metabolites. Calculations as in Table 8.3. All quantities are given in $kJmol^{-1}$.

Metabolite	$\Delta_f G$	$\Delta_r G$ (ΔG_{acid})	$\Delta_r G$ lit	$\Delta_f H$	$\Delta_r H$ (ΔH_{acid})	$\Delta_r H$ lit
Aminoacids						
Alanine	-411	1400	1400[a], 1398.5[c]	-524	1431	1432[a,b] 1425±8.8
Arginine	-273	1356	1347[a], 1359.5[c]	-532	1388	1381[a], 1387[b] 1389±13
Asparagine	-572	1353	1354[a], 1359[c]	-760	1385	1386[a], 1384[b] 1388±13

Aspartic acid 1 Aspartate*	-830 -778	1305 1356	1316[b] 1364[a]	-991 -938	1332 1385	1349[a], 1345[b] 1394[a]
Cysteine	-411	1374	1369[a] 1364[c]	-525	1404	1399[a], 1396[b] 1393±13
Glutamic acid Glutamate*	-812 -795	1332 1349	1324[a] 1357[a]	-997 -979	1355 1373	1347[a], 1349[b] 1384[a]
Glutamine	-558	1351	1347[a], 1359[c]	-777	1382	1378[a,b] 1388±13
Glycine	-409	1404	1403[a], 1402[c]	-491	1434	1435[a], 1434[b] 1433[d]
Histidine	-252	1348	1345[a], 1356[c]	-447	1377	1376[a], 1374[b] 1385±13
Isoleucine	-394	1399	1396[a], 1388.5[c]	-601	1426	1426[a,b] 1418±13
Leucine	-395	1398	1395[a], 1390[c]	-602	1427	1424[a], 1428[b] 1419±13
Lysine	-329	1390	1380[a] 1383[c]	-576	1416	1410[a], 1415[b] 1412±13
Methionine	-356	1403	1385[a], 1376[c]	-534	1427	1418[a], 1412[b] 1405±13
Phenylalanine	-268	1388	1384[a], 1379[c]	-433	1424	1416[a], 1417[b] 1408±13
Proline	-341	1394	1394[a] 1395[c]	-499	1425	1425[a], 1430[b] 1430±13
Serine	-583	1363	1363[a,c]	-722	1393	1392[a,b] 1392±13
Threonine	-592	1360	1359[a], 1360.5[c]	-763	1390	1390[a], 1397[b] 1390±13
Tryptophan	-155	1388	1390[a], 1380.5[c]	-358	1422	1423[a], 1422[b] 1410±13
Tyrosine	-415	1386	1382[a] 1378.5[c]	-602	1420	1415[a], 1419[b] 1408±13
Valine	-398	1391	1394[a], 1391[c]	-577	1427	1425[a], 1430[b] 1420±13

[a]Stover 2012[157] ; [b]Jones et al 2007[159] ;[c]O'Hair 1995[150], [d]Bouchoux 2011[158].
*Deprotonated side carboxyl.

In Table 7.5 we report the gas-phase *basicities* and proton affinities for aminoacids, adenosine and βD-Glucose as well as the standard enthalpies and Gibbs free energies of formation of the protonated metabolites. The latter quantities were obtained from the GBs and PAs recommended in the literature and the corresponding quantities for the neutral metabolites from Table 7.3. We observe that our calculations conform with

literature data in general. The obtained formation quantities are in most cases positive but no obvious conclusions can be drawn out of their values.

Table 7-5: Estimated standard thermochemical quantities (ideal gas phase) of ionic metabolites from corresponding properties for neutral metabolites (Table 7.3) and their recommended *basicities* (GB) and proton affinities (PA)a. Values in parenthesis are our GB and PA calculations at the same level of theory as in Table 4. All quantities are given in $kJmol^{-1}$.

Metabolite	**GB[a]**	**$\Delta_f G^0$**	**PA[a]**	**$\Delta_f H^0$**
Alanine	868 (872)	347.0	902 (904)	209.0
Arginine Arginine 2*	1007 (954)	389.8 429.6	1046 (985)	99.7 149.1
Asparagine	905	196.6	942	-21.0
Aspartic acid	882	9.2	920	-177.2
Cysteine	870 (874)	370.2	903 (905)	233.5
Glutamic acid	908	-26.0	947	-233.0
Glutamine	935	182.2	975	-68.0
Glycine	854 (854)	358.9	887 (887)	254.3
Histidine	947	475.5	979	264.1
Isoleucine	885 (884)	348.3	919 (921)	120.5
Leucine	883 (885)	349.4	916 (920)	121.8
Lysine	952 (948)	355.1	994 (990)	80.2
Methionine	899 (910)	366.2	938 (945)	167.5
Phenylalanine	892 (896)	478.9	930 (929)	279.5
Proline	908	383.2	942	200.0
Serine	878 (876)	201.5 205.5	912 (908)	39.5
Threonine	886 (885)	187.5	919 (917)	-6.0
Tryptophan	909	574.2	945	341.1
Tyrosine	895	329.6	933	111.0
Valine	881 (887)	349.8	915 (920)	145.4
Adenosine	945[b]	445	979[b]	72.4
βD-Glucose	786.6[c]	-111.1	810.3[c]	-392.2

[a]Bouchoux 2012[158], [b]Bouchoux 2008[156], [c]Jebber 1996[228]. *Main chain –NH2

Appendix Table A4 summarizes the derived formation quantities, together with some formation quantities from the literature and shows a good consistency between our calculations and literature values. The formation quantities for all three types of species (neutral, protonated, deprotonated) follow similar trends. This is better visualized in figures 7.5 and figure 7.6. Figure 7.5 shows that the free energies of formation of protonated and deprotonated aminoacids vary *linearly* with the free energy of formation of the neutral aminoacids. The parameters of the straight lines are given in Appendix Table A5. The two lines in figure 7.5 are almost parallel with a slope close to one. The deviations from the straight line are mainly due to aminoacids having strong protonation sites other than the common -NH2 group in the main chain. As expected, the best example is arginine (Arg1 in figure 7.5) with the strong basic guanidine site. We also indicate in figure the free energy of formation of arginine protonated on the common –NH2 group (Arg2 in figure 7.5), which is now closer to the linear approximation.

Similar comments can be made for the enthalpy of formation of protonated and deprotonated aminoacids (figure 7.6).

Figures 7.5 and 7.6 are useful for rationalizing *acidities* and *basicities* of aminoacids and their derivatives. It is also worth mentioning that one of the other protonated species (adenosine) has its formation quantities close to the straight lines in these figures. The other protonated species (βD-glucose), however, deviates from the lines to some extent. This diverse behavior requires some explanation.

Equation 8.7 may be rewritten as follows:

$$GB\ (M) = \Delta_f G^0(M) + \Delta_f G^0(H^+) - \Delta_f G^0(MH^+) \qquad (7.7a)$$

Or

$$\Delta_f G^0(MH^+) = \Delta_f G^0(M) + \Delta_f G^0(H^+) - GB\ (M) \qquad (7.11)$$

If R is the reference compound of the isodesmic reaction, which is often used to replace for the $\Delta_f G^0(H^+)$, equation (8.11) can be rewritten as:

$$\Delta_f G^0(RH^+) = \Delta_f G^0(R) + \Delta_f G^0(H^+) - GB\ (R) \qquad (7.12)$$

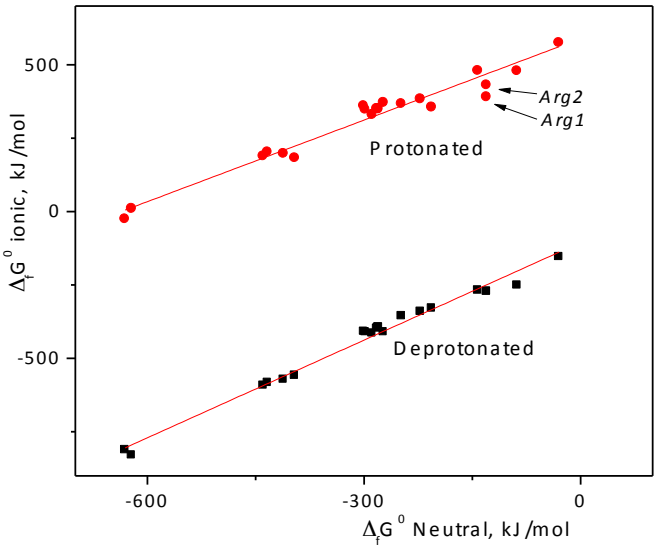


Figure 7-5: Formation free energy of protonated and deprotonated *aminoacids* vs. the formation free energy of their neutral counterparts

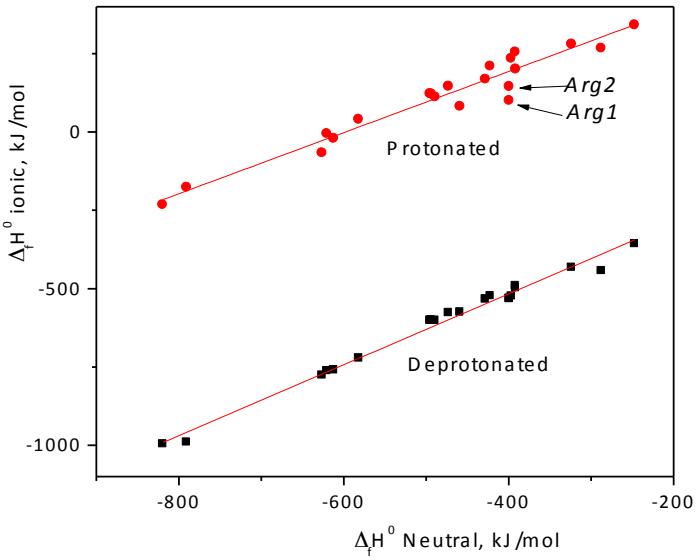


Figure 7-6: Formation enthalpies of protonated and deprotonated *aminoacids* vs. the formation enthalpy of their neutral counterparts.

Replacing in equation (7.11), we obtain:

$$\Delta_f G^0(MH^+) = \Delta_f G^0(M) + \Delta_f G^0(RH^+) - \Delta_f G^0(R) + GB\ (R) - GB\ (M) \qquad (7.13)$$

A linear relationship (figure 7.5) in the form of

$\Delta_f G^0(MH^+) = s\Delta_f G^0(M) + I = s\Delta_f G^0(M) + [\Delta_f G^0(RH^+) - s\Delta_f G^0(R)]$

or

$$\Delta_f G^0(MH^+) = \Delta_f G^0(RH^+) + s[\Delta_f G^0(M) - \Delta_f G^0(R)] \quad (7.14)$$

implies that:

$$s\Delta_f G^0(M) = \Delta_f G^0(M) + s\Delta_f G^0(R) - \Delta_f G^0(R) + GB\,(R) - GB\,(M) \quad (7.15)$$

which can be restated as:

$$(s-1)\Delta_f G^0(M) = (s-1)\Delta_f G^0(R) + GB\,(R) - GB\,(M) \quad (7.16)$$

i.e.,

$$GB\,(M) = GB\,(R) + (s-1)[\Delta_f G^0(R) - \Delta_f G^0(M)] \quad (7.17)$$

Equation 7.14 or equivalently equation 7.17 implies that the constant (intercept) term of the above linear relationship is dictated by the corresponding thermochemical quantities of the reference compound. Amines are often used to anchor gas phase basicities of aminoacids while for saccharides a more appropriate reference compound would be an alcohol. Proton affinities and gas-phase acidities can be treated in a similar way.

Figures 7.7 and 7.8 are analogous to figures 7.5 and 7.6 with the difference of enhanced scales in order to accommodate the reference compounds as well. Parameters and statistics of the linear fittings are shown in Appendix Table A6. As can be observed, there is now much better linear fit in all cases. Classes of compounds anchored to a given reference fall on a straight line passing through reference compounds. Sugars, as an example, fall on the lines passing through alcohols. Aminoacids, amines, and nucleobases fall on nearly the same lines. The lines in each graph are almost parallel. This feature is particularly useful for a quick qualitative estimation of the formation quantities of ions or, equivalently, the gas-phase acidities and basicities from the formation quantities of neutral counterparts. We emphasize the point that these lines are for a qualitative estimation of gas phase acidities and basicities. As pointed out

above, the very same neutral metabolite may exhibit different acid/base strength depending on its protonation/deprotonation sites. Thus, the observed scatter in the figures was expected. Nevertheless, the straight lines of figures 7.7 and 7.8 are useful tools in discussing gas-phase acidities and basicities.

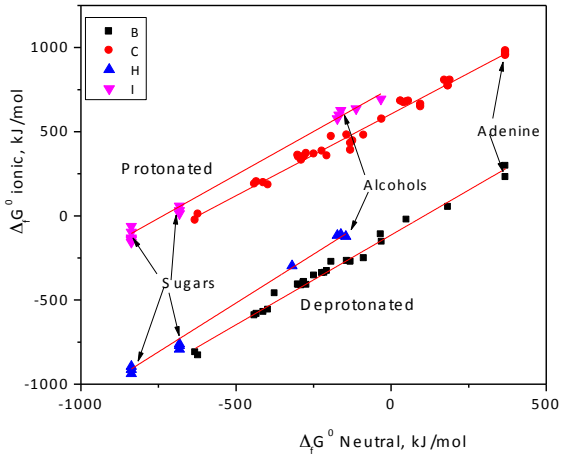


Figure 7-7: Formation free energies of protonated and deprotonated compounds (including aminoacids) vs. the formation enthalpy of their neutral counterparts. The lines are linear fits (parameters and statistics are shown in Appendix Table A5).

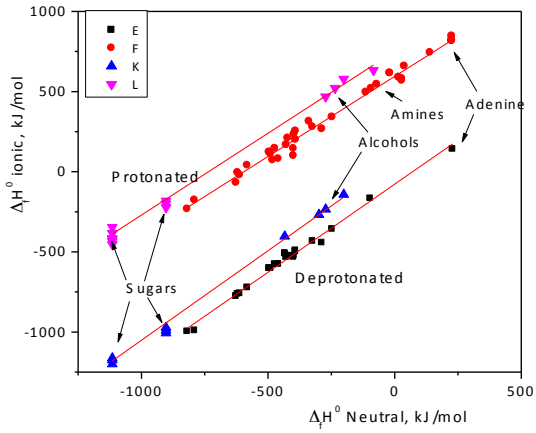


Figure 7-8: Formation enthalpies of protonated and deprotonated compounds vs. the formation enthalpy of their neutral counterparts. The lines are linear fits (parameters and statistics are shown in Appendix Table A6).

7.4 Conclusions

The primary objective of the present work was the reliable estimation of the heats and Gibbs free energies of formation in the ideal gas state for the key metabolites, such as the aminoacids, oligosaccharides and nucleosides. The adopted route of quantum chemical calculations was through the study of the appropriate isodesmic reactions of metabolites. These isodesmic reactions are based on the thermochemical data for the other reactants and products and therefore, we used the most reliable values for these thermochemical quantities of the auxiliary reactants and products. This leads to a sufficiently reliable set of predicted standard thermochemical quantities of the studied metabolites. Despite the scarcity of available reliable experimental data and the dispersion of available theoretical calculations, our calculations are consistent with the more recent and reliable quantum thermochemical calculations in literature. In the case of aminoacids, thermochemical quantities were also estimated for their protonated and deprotonated ions. Furthermore, we provided a first comprehensive thermochemical table together with the accompanying figures encompassing neutral, protonated and deprotonated aminoacids, βD-glucose, βD-ribose and adenosine including all 20 aminoacids. The observed linear trends may have broader implications that deserve further systematic studies. Calculations of the present work will be very useful in subsequent GCM and hydration works.

Chapter 8

CONCLUSIONS & PRESPECTIVES

Cellular modification for designing reliable synthetic microbial cell factories (SMCFs) lies at the heart of metabolic engineering. This is a multidiscipline process that entails numerous challenges and cannot be carried out intuitively, which necessitates the use of several computational methods and tools.

With the work we put forward in this thesis, we outlined the significance of several computational disciplines such as computational biochemistry, bioinformatics, computational biology, cheminformatics, constraint-based modeling, process engineering, mathematical optimization, databases, etc. for the profitability and the progress of systems biology and in particular the metabolic engineering studies.

After introducing several principal milestones of computational metabolic engineering concepts in **Chapetr 2**, we described the details of our developed methodologies and their potential biochemical and biomedical applications and they could contribute in gaining novel biological and biochemical insights.

We presented in **Chapter 3** our results from the computational investigation of the entire space of enzymatic/metabolic possibilities for the design of *de novo* metabolic pathways, irrespective of the choice of the organism. We discovered that around 120'000 new biochemical/enzymatic reactions could evolve from the known biochemistry. We further compared the structural similarities of the hypothetical novel reactions with those already existing in the metabolic databases and cataloged our findings in a web-based tool.

In **Chapter 4**, due to the importance of lipids in cellular physiology and pathology as well as recent interest for their role as resources for alternative fuels and chemicals, we had a closer look to lipid metabolism and applied several computational methods to get new insights about the structural diversities and metabolic functions of this

metabolism. We introduced the computational pipeline NICELips used for elaborating our knowledge on lipid metabolism. The results of our case studies prove the effectiveness of NICELips for integrating the lipid species having unknown metabolism (obtained in lipidomics studies) into metabolic pathways. Broadening the application of NICELips to all the subsections of lipid metabolism will result in the establishment of a new repository of lipid information with potential applications in the design of new therapeutical approaches for lipid related disorders.

In **Chapter 5** we investigated the metabolic network of *E.coli* one of the most well studied chassis microorganisms. Enormous effort has been spent to discover all the metabolic potential of this organism for the microbial production of native and non-native compounds. While several ongoing research studies focus on all fronts to discover new insights, a computational strategy to discover the full theoretical enzymatic potential of this or any given organism was missing.

In addressing so, we developed "super *E. coli*", an artificial reconstruction of the *E. coli* metabolic network with all possible new functionalities and extra metabolic capabilities, discovered on the basis of its known biochemistry. The metabolic network of "super *E. coli*" fulfills major objectives such as increasing the yield through biomass production and in addition, encompassing the synthetic capacity for the biosynthesis of several heterologous compounds. We can envision that our approach will provide the blueprints of new organisms able to efficiently convert low value compounds into every possible specialty and commodity chemical.

In **Chapter 6** we took one step beyond the conventional analyses of metabolism, in which the metabolic network is represented as a graph of metabolites interconnected with biological reactions. We developed the computational framework "iAM.NICE" for complementing the metabolic networks by accounting for all atom transitions from the substrates to the products. We demonstrated our results through the reconstruction of an atom-level representation of the *E. coli* core metabolism and we elucidated the mass flow through the network. The results of this study and the further application of "iAM.NICE" for the atom-level reconstruction of other biological networks rather than those discussed in this work could provide detailed predictions for guiding the ^{13}C Metabolic Flux Analysis. We believe that the ability to accurately track the moments of atoms in the metabolic networks opens up new opportunities for revealing hidden parts of metabolism and for generating hypotheses regarding the alternative or missing

pathways in metabolism with high potential applications in metabolic engineering studies.

Metabolism centers on the bioenergetics of cellular processes, and in our last research project of this thesis in **Chapter 7**, we focused on the thermodynamics of metabolism in order to address the existing constraints and limitations for obtaining/estimating the thermodynamics properties of metabolite and consequently the metabolic reactions. We proposed a sound pipeline for the accurate estimation of several thermodynamic properties for a broad range of metabolites. Furthermore, we proposed a new approach for adjusting the measured/calculated/estimated thermodynamics properties to any temperature and pressure, rather than the standard ones which do not hold true especially in biological systems. In this work, we focused our attention on the estimation of thermodynamic properties of heavy metabolites in the gas phase. In the future, applying the concept of the thermodynamic cycles coupled with the data obtained in this thesis, we can provide estimated thermodynamic values for the metabolites in aqueous solutions (the same as for biological systems), which are adjustable to relevant temperature, and pressure conditions based on the nature of the study.

We close this thesis by highlighting the fundamental position of the computational tools and approaches in systems biology and in the success of its pioneer research field, metabolic engineering. During this PhD work, we contributed to the field of computational systems biology through developing new tools and methods for enhancing our understanding of metabolic networks. This work can lead to the discovery of new insights, the generation of new hypotheses and the design of novel strategies toward the ultimate goal of metabolic engineering: *to be able to produce any molecule from any renewable resource, using SMCFs*.

APPENDIX

Table A1: Proposed EC classification number up to the third level by BNICE.ch for 178 KEGG reactions that are missing an EC number. The three columns show the KEGG reaction identifier, the reaction rule(s) catalyzing the reaction and the suggested EC number for each reaction. Reactions for which two or more EC numbers are suggested are colored in light red.

KEGG ID	Reaction rules	Suggested EC number
R00091	|1.8.1B1(rev)|1.8.1B1	1.8.1.-
R00270	|3.5.1A1(rev)|3.5.1A1|6.3.1A1(rev)|6.3.1A1	3.5.1.- | 6.3.1.-
R00598	|2.1.1A5(rev)|2.1.1A5|2.1.1A1(rev)	2.1.1.-
R00683	|2.1.1A5(rev)|2.1.1A5	2.1.1.-
R00744	|1.2.1A2(rev)|1.2.1A2	1.2.1.-
R01339	|2.6.1A1(rev)	2.6.1.-
R02343	|1.1.1A1(rev)|1.1.1A1|1.2.1B1(rev)|1.2.1B1	1.1.1.- | 1.2.1.-
R02344	|1.1.1A1(rev)|1.1.1A1|1.2.1B1(rev)|1.2.1B1	1.1.1.- | 1.2.1.-
R02583	|4.2.1A1(rev)|4.2.1A1|4.2.1A2(rev)|4.2.1A2|4.2.1A6(rev)|4.2.1A6	4.2.1.-
R02659	|3.1.1A1(rev)|3.1.1A1	3.1.1.-
R02798	|4.2.1A2(rev)|4.2.1A2|4.2.1A6(rev)|4.2.1A6	4.2.1.-
R02801	|4.2.1A1(rev)|4.2.1A1|4.2.1A2(rev)|4.2.1A2|4.2.1A6(rev)|4.2.1A6	4.2.1.-
R02856	|1.1.1A2(rev)|1.1.1A2|1.2.1B2(rev)|1.2.1B2	1.1.1.- | 1.2.1.-
R03082	|1.2.1D3(rev)	1.2.1.-
R03128	|5.3.2A1(rev)|5.3.2A1	5.3.2.- | 5.3.2.-
R03249	|1.5.1A1(rev)|1.5.1A1	1.5.1.- | 1.5.1.-
R03250	|3.5.3A1(rev)|3.5.3A1	3.5.3.- | 3.5.3.-
R03314	|1.5.1A1(rev)|1.5.1A1	1.5.1.- | 1.5.1.-
R03383	|3.1.2A1(rev)|3.1.2A1|3.3.1A1|6.2.1A1(rev)|6.2.1A1|6.2.1B1(rev)|6.2.1B1|6.2.1C1(rev)|6.2.1C1|6.2.1D1(rev)|6.2.1D1	3.1.2.- | 3.3.1.- | 6.2.1.-
R03506	|1.1.1D1(rev)|1.1.1D1|1.2.1A1(rev)|1.2.1A1	1.1.1.- | 1.2.1.-
R03507	|1.1.1A1(rev)|1.1.1A1|1.2.1B1(rev)|1.2.1B1	1.1.1.- | 1.2.1.-
R03694	|4.2.1A1(rev)|4.2.1A1|4.2.1A2(rev)|4.2.1A2|4.2.1A6(rev)|4.2.1A6|4.2.1C1(rev)|4.2.1C1	4.2.1.-
R03695	|3.1.2A1(rev)|3.1.2A1|3.3.1A1|6.2.1A1(rev)|6.2.1A1|6.2.1B1(rev)|6.2.1B1|6.2.1C1(rev)|6.2.1C1|6.2.1D1(rev)|6.2.1D1	3.1.2.- | 3.3.1.- | 6.2.1.-
R03758	|1.1.1B2(rev)|1.1.1B2|4.1.1A2(rev)|4.1.1A2|4.1.1A3(rev)|4.1.1A3|4.1.1A6(rev)|4.1.1A6|4.1.1A7(rev)|4.1.1A7|4.2.1A4(rev)	1.1.1.- | 4.1.1.- | 4.2.1.-
R03863	|1.1.1A1(rev)|1.1.1A1|1.2.1B1(rev)|1.2.1B1	1.1.1.- | 1.2.1.-
R03864	|1.1.1A2(rev)|1.1.1A2|1.2.1B2(rev)|1.2.1B2	1.1.1.- | 1.2.1.-
R03892	|5.3.2A1(rev)|5.3.2A1|5.3.2A2(rev)|5.3.2A2	5.3.2.-
R03897	|4.2.1A1(rev)|4.2.1A1|4.2.1A2(rev)|4.2.1A2|4.2.1A6(rev)|4.2.1A6|4.2.1C1(rev)|4.2.1C1	4.2.1.-
R04269	|2.6.1A1(rev)	2.6.1.-
R04416	|4.2.1A2(rev)|4.2.1A2|4.2.1A6(rev)|4.2.1A6|4.2.1C1(rev)|4.2.1C1	4.2.1.-
R04443	|1.5.1A1(rev)|1.5.1A1	1.5.1.-
R04446	|1.5.1A1(rev)|1.5.1A1	1.5.1.-

R04453	|1.2.1A2(rev)|1.2.1A2	1.2.1.-
R04593	|3.1.1A1	3.1.1.-
R04646	|2.4.1D1(rev)|2.4.1A1(rev)|2.4.1B1(rev)|2.4.1C1(rev)	2.4.1.-
R04875	|3.5.1A1	3.5.1.-
R04876	|1.1.1A1(rev)|1.2.1B1(rev)	1.1.1.- | 1.2.1.-
R04918	|4.1.1A1(rev)|4.1.1A1|4.1.1A6(rev)|4.1.1A6|4.1.1A7(rev)|4.1.1A7|4.2.1A4(rev)	4.1.1.- | 4.2.1.-
R04987	|1.14.13A1(rev)|	1.14.13.-
R04999	|3.1.1A1	3.1.1.-
R05056	|5.3.3A1(rev)|5.3.3A1|5.3.3A2(rev)|5.3.3A2	5.3.3.-
R05057	|5.3.3A1(rev)|5.3.3A1|5.3.3A2(rev)|5.3.3A2|5.3.3A3(rev)|5.3.3A3	5.3.3.-
R05058	|5.3.3A1(rev)|5.3.3A1|5.3.3A2(rev)|5.3.3A2	5.3.3.-
R05080	|2.4.1E1(rev)|2.4.1A1(rev)|2.4.1B1(rev)|2.4.1C1(rev)	2.4.1.-
R05107	|2.4.1A1|2.4.1E1(rev)|2.4.1A1(rev)|2.4.1B1(rev)|2.4.1C1(rev)	2.4.1.-
R05108	|2.4.1C1|2.4.1E1(rev)|2.4.1A1(rev)|2.4.1B1(rev)|2.4.1C1(rev)	2.4.1.-
R05119	|1.2.1A1(rev)|1.2.1A1	1.2.1.-
R05125	|5.3.3A1(rev)|5.3.3A1|5.3.3A2(rev)|5.3.3A2	5.3.3.-
R05232	|1.1.1A1(rev)|1.1.1A1|1.2.1B1(rev)|1.2.1B1	1.1.1.- | 1.2.1.-
R05250	|1.14.13A1(rev)|	1.14.13.-
R05303	|2.1.1A5(rev)|2.1.1A5|2.1.1A1(rev)	2.1.1.-
R05445	|3.8.1A2(rev)|3.8.1A2	3.8.1.-
R05471	|1.1.1A2(rev)|1.1.1A2|1.2.1B2(rev)|1.2.1B2	1.1.1.- | 1.2.1.-
R05477	|3.8.1A2(rev)|3.8.1A2	3.8.1.-
R05478	|4.2.1A2(rev)|4.2.1A2|4.2.1A6(rev)|4.2.1A6|4.2.1C1(rev)|4.2.1C1	4.2.1.-
R05515	|1.1.1A1(rev)|1.1.1A1|1.2.1B1(rev)|1.2.1B1	1.1.1.- | 1.2.1.-
R05535	|4.5.1A1(rev)|4.5.1A1	4.5.1.-
R05543	|4.2.1A2|4.2.1A6	4.2.1.-
R05569	|3.5.99A1(rev)|3.5.99A1	3.5.99.-
R05609	|5.3.3A1(rev)|5.3.3A1|5.3.3A2(rev)|5.3.3A2	5.3.3.-
R05656	|5.5.1A5(rev)|5.5.1A5	5.5.1.-
R05659	|4.2.1A1(rev)|4.2.1A1|4.2.1A2(rev)|4.2.1A2|4.2.1A6(rev)|4.2.1A6	4.2.1.-
R05677	|3.5.1A1|6.3.1A1(rev)	3.5.1.- | 6.3.1.-
R05694	|1.1.1A2(rev)|1.1.1A2|1.2.1B2(rev)|1.2.1B2	1.1.1.- | 1.2.1.-
R05856	|4.2.1A6(rev)|4.2.1A6	4.2.1.-
R05866	|1.14.13A5(rev)|1.14.13A5	1.14.13.-
R05870	|2.3.1F1(rev)|2.3.1D1|2.3.1D1(rev)	2.3.1.-
R05881	|2.1.1A4(rev)|2.1.1A5(rev)|2.1.1A5|2.1.1A1(rev)	2.1.1.-
R06418	|5.3.3A1(rev)|5.3.3A1|5.3.3A2(rev)|5.3.3A2|5.3.3A3(rev)|5.3.3A3	5.3.3.-
R06428	|4.2.1A1(rev)|4.2.1A1|4.2.1A2(rev)|4.2.1A2|4.2.1A6(rev)|4.2.1A6	4.2.1.-
R06429	|1.3.1A2(rev)|1.3.1A2	1.3.1.-
R06430	|1.3.1A2(rev)|1.3.1A2	1.3.1.-
R06432	|2.1.1A1|2.1.1A1(rev)	2.1.1.-
R06434	|1.1.1A2(rev)|1.1.1A2|1.2.1B2(rev)|1.2.1B2	1.1.1.- | 1.2.1.-
R06472	|1.1.1A2(rev)|1.1.1A2|1.2.1B2(rev)|1.2.1B2	1.1.1.- | 1.2.1.-
R06473	|1.1.1A2(rev)|1.1.1A2|1.2.1B2(rev)|1.2.1B2	1.1.1.- | 1.2.1.-

R06474	|1.1.1A2(rev)|1.1.1A2|1.2.1B2(rev)|1.2.1B2	1.1.1.- | 1.2.1.-
R06475	|1.1.1A2(rev)|1.1.1A2|1.2.1B2(rev)|1.2.1B2	1.1.1.- | 1.2.1.-
R06592	|4.2.3A2(rev)|4.2.3A2	4.2.3.-
R06593	|4.2.1A1(rev)|4.2.1A1|4.2.1A6(rev)|4.2.1A6	4.2.1.-
R06594	|1.1.1A1(rev)|1.1.1A1|1.2.1B1(rev)|1.2.1B1	1.1.1.- | 1.2.1.-
R06627	|2.1.1A4(rev)|2.1.1A5(rev)|2.1.1A5|2.1.1A1(rev)	2.1.1.-
R06629	|1.1.1A2(rev)|1.1.1A2|1.2.1B2(rev)|1.2.1B2	1.1.1.- | 1.2.1.-
R06665	|4.2.1A2(rev)|4.2.1A2|4.2.1A6(rev)|4.2.1A6	4.2.1.-
R06685	|3.1.1A1	3.1.1.-
R06686	|4.1.1A6(rev)|4.1.1A6|4.1.1A7(rev)|4.1.1A7	4.1.1.-
R06752	|1.14.15A1(rev)|1.14.15A1	1.14.15.-
R06774	|2.1.1A5(rev)|2.1.1A5|2.1.1A1(rev)	2.1.1.-
R06830	|1.1.1B2(rev)|1.1.1B2|4.1.1A2(rev)|4.1.1A2|4.1.1A3(rev)|4.1.1A3|4.1.1A6(rev)|4.1.1A6|4.1.1A7(rev)|4.1.1A7|4.2.1A4(rev)	1.1.1.- | 4.1.1.- | 4.2.1.-
R06866	|4.1.1A6(rev)|4.1.1A6|4.1.1A7(rev)|4.1.1A7	4.1.1.-
R06867	|2.4.1E1(rev)|2.4.1A1(rev)|2.4.1B1(rev)|2.4.1C1(rev)	2.4.1.-
R06968	|3.5.99A1(rev)|3.5.99A1	3.5.99.-
R06997	|2.3.1F1(rev)|2.3.1D1|2.3.1D1(rev)	2.3.1.-
R07009	|5.5.1A5(rev)|5.5.1A5	5.5.1.-
R07010	|5.5.1A5(rev)|5.5.1A5	5.5.1.-
R07011	|5.5.1A5(rev)|5.5.1A5	5.5.1.-
R07012	|5.5.1A5(rev)|5.5.1A5	5.5.1.-
R07016	|1.14.13A5(rev)|1.14.13A5	1.14.13.-
R07058	|1.1.1A1(rev)|1.1.1A1|1.2.1B1(rev)|1.2.1B1	1.1.1.- | 1.2.1.-
R07067	|4.2.1A2(rev)|4.2.1A2|4.2.1A6(rev)|4.2.1A6|4.2.1C1(rev)|4.2.1C1	4.2.1.-
R07073	|5.5.1A5(rev)|5.5.1A5	5.5.1.-
R07074	|5.5.1A5	5.5.1.-
R07076	|5.5.1A5	5.5.1.-
R07086	|5.5.1A5(rev)|5.5.1A5	5.5.1.-
R07095	|4.2.1A6	4.2.1.-
R07101	|4.2.1A6	4.2.1.-
R07124	|3.1.2A1|3.3.1A1|6.2.1A1(rev)|6.2.1B1(rev)|6.2.1C1(rev)|6.2.1D1(rev)	3.1.2.- | 6.2.1.-
R07427	|5.3.3A3(rev)|5.3.3A3	5.3.3.-
R07428	|1.1.1A1(rev)|1.1.1A1|1.2.1B1(rev)|1.2.1B1	1.1.1.- | 1.2.1.-
R07474	|1.14.15A1(rev)|1.14.15A1	1.14.15.-
R07497	|5.3.3A3(rev)|5.3.3A3	5.3.3.-
R07506	|1.3.1A2(rev)|1.3.1A2	1.3.1.-
R07576	|4.2.1A1	4.2.1.-
R07707	|5.5.1A5(rev)|5.5.1A5	5.5.1.-
R07708	|4.2.1A2(rev)|4.2.1A2|4.2.1A6(rev)|4.2.1A6	4.2.1.-
R07922	|2.1.1A4(rev)|2.1.1A5(rev)|2.1.1A5|2.1.1A1(rev)	2.1.1.-
R07986	|4.2.1A1(rev)|4.2.1A1|4.2.1A2(rev)|4.2.1A2|4.2.1A6(rev)|4.2.1A6|4.2.1C1(rev)|4.2.1C1	4.2.1.-
R07991	|4.2.1A1	4.2.1.-
R07992	|4.2.1A1	4.2.1.-
R08024	|4.2.1A1	4.2.1.-

R08061	|3.6.1A1(rev)|3.6.1A1	3.6.1.-
R08062	|3.6.1A1(rev)|3.6.1A1	3.6.1.-
R08063	|3.6.1A1(rev)|3.6.1A1	3.6.1.-
R08064	|3.6.1A1(rev)|3.6.1A1	3.6.1.-
R08069	|3.1.3A1(rev)|3.1.3A1	3.1.3.-
R08070	|3.1.3A1(rev)|3.1.3A1	3.1.3.-
R08071	|3.1.3A1(rev)|3.1.3A1	3.1.3.-
R08079	|3.1.3A1(rev)|3.1.3A1	3.1.3.-
R08128	|2.4.1A1|2.4.1E1(rev)|2.4.1A1(rev)|2.4.1B1(rev)|2.4.1C1(rev)	2.4.1.-
R08145	|6.3.5A1(rev)|6.3.5A1	6.3.5.-
R08148	|1.14.13A5(rev)|1.14.13A5	1.14.13.-
R08152	|3.3.2A1(rev)|3.3.2A1	3.3.2.-
R08153	|3.3.2A1(rev)|3.3.2A1	3.3.2.-
R08154	|2.7.1A1(rev)|2.7.1A1|2.7.8A1	2.7.1.- | 2.7.8.-
R08251	|3.5.1A1|6.3.1A1(rev)	3.5.1.- | 6.3.1.-
R08279	|4.2.1A1	4.2.1.-
R08289	|4.2.1A1	4.2.1.-
R08309	|4.3.2A1	4.3.2.-
R08311	|3.5.1A1	3.5.1.-
R08328	|4.1.1A1(rev)|4.1.1A1|4.1.1A6(rev)|4.1.1A6|4.1.1A7(rev)|4.1.1A7|4.2.1A4(rev)	4.1.1.- | 4.2.1.-
R08451	|3.5.1A1	3.5.1.-
R08492	|1.1.1A2(rev)|1.1.1A2|1.2.1B2(rev)|1.2.1B2	1.1.1.- | 1.2.1.-
R08494	|1.1.1A2(rev)|1.1.1A2|1.2.1B2(rev)|1.2.1B2	1.1.1.- | 1.2.1.-
R08561	|1.1.1A2(rev)|1.1.1A2|1.2.1B2(rev)|1.2.1B2	1.1.1.- | 1.2.1.-
R08562	|1.1.1A2(rev)|1.1.1A2|1.2.1B2(rev)|1.2.1B2	1.1.1.- | 1.2.1.-
R08563	|1.1.1A2(rev)|1.1.1A2|1.2.1B2(rev)|1.2.1B2	1.1.1.- | 1.2.1.-
R08590	|1.1.1A1(rev)|1.2.1B1(rev)	1.1.1.- | 1.2.1.-
R08591	|5.3.2A1(rev)|5.3.2A1|5.3.2A2(rev)|5.3.2A2	5.3.2.-
R08861	|1.1.1A2(rev)|1.1.1A2|1.2.1B2(rev)|1.2.1B2	1.1.1.- | 1.2.1.-
R08870	|2.3.3A1(rev)|2.3.3A1|2.3.3B1(rev)|2.3.3B1	2.3.3.-
R08871	|2.3.1D3(rev)|2.3.1F1(rev)	2.3.1.-
R08874	|3.5.1A1|6.3.1A1(rev)	3.5.1.- | 6.3.1.-
R08876	|3.5.1A1|6.3.1A1(rev)	3.5.1.- | 6.3.1.-
R09170	|5.5.1A5(rev)|5.5.1A5	5.5.1.-
R09172	|5.5.1A5(rev)|5.5.1A5	5.5.1.-
R09175	|5.5.1A5(rev)|5.5.1A5	5.5.1.-
R09179	|5.5.1A5	5.5.1.-
R09253	|4.3.1A3(rev)|4.3.1A3	4.3.1.-
R09259	|4.1.2A2	4.1.2.-
R09260	|4.1.2A2	4.1.2.-
R09261	|4.1.2A2	4.1.2.-
R09273	|3.1.1A1	3.1.1.-
R09276	|1.13.11A1(rev)|1.13.11A2(rev)|1.13.11A2	1.13.11.-
R09278	|4.1.1A6(rev)|4.1.1A6|4.1.1A7(rev)|4.1.1A7	4.1.1.-

R09282	|4.2.1A1(rev)|4.2.1A1|4.2.1A6(rev)|4.2.1A6|4.2.1C1(rev)|4.2.1C1	4.2.1.-
R09312	|1.6.5A1(rev)|1.6.5A1	1.6.5.-
R09337	|3.1.1A1	3.1.1.-
R09342	|4.2.1A1(rev)|4.2.1A1|4.2.1A6(rev)|4.2.1A6	4.2.1.-
R09377	|3.5.4A1(rev)|3.5.4A1	3.5.4.-
R09779	|3.5.1A1|6.3.1A1(rev)|6.3.2A1(rev)	3.5.1.- | 6.3.1.- | 6.3.2.-
R09850	|2.7.4B1	2.7.4.-
R09854	|2.3.1F1(rev)	2.3.1.-
R10013	1.5.1A1|	1.5.1.-
R10212	1.2.3B1|	1.2.3.-
R10458	2.1.1A5|	2.1.1.-
R10459	1.1.1A1(rev)|1.2.1B1(rev)|	1.1.1.- | 1.2.1.-
R10638	2.4.1A1|	2.4.1.-
R10639	2.4.1A1|	2.4.1.-

Table A2: List of 81 reactions that BNICE.ch proposed them as “novel reactions” based on theKEGG 2012 database and they are reported as “known reactions” in KEGG 2014 database. It shows the capability of BNICE.ch for prediction reactions that follow known biochemistry rules.

KEGG reaction	Equation	EC number	Reaction rule
R10013	C20279<=>C00001+C16138		1.5.1A1|
R10021	C00003+C03448<=>C00004+C00080+C03037	1.1.1.327	1.1.1A1|1.2.1B1|
R10037	C00029+C18796<=>C00015+C08334	2.4.1.63	2.4.1A1|
R10042	C00004+C00007+C00080+C01407<=>C00001+C00003+C00146	1.14.13.0	1.14.13A2|
R10043	C00004+C00007+C00080+C00146<=>C00001+C00003+C00090	1.14.13.0	1.14.13A2|
R10044	C00005+C00007+C00080+C07160<=>C00001+C00006+C20321	1.14.13.160	1.14.13A7|
R10046	C00001+C05841<=>C00121+C00253	3.2.2.3	3.3.1A1|
R10051	C00001+C20320<=>C20324	3.7.1.18	3.7.1A1|
R10052	C00003+C04411<=>C00004+C00011+C00080+C00233	1.1.1.85	1.1.1.A9|
R10075	C00006+C20325<=>C00005+C00080+C20308	1.3.1.92	1.3.1A2|
R10078	C20329<=>C00001+C11887	1.14.13.109	4.2.1A6|
R10084	C00001+C20331<=>C20332	1.14.99.36	3.3.2A1|
R10093	C00002+C20345<=>C00008+C00129	2.7.4.26	2.7.4A1|2.7.4B1|
R10094	C00026+C20350<=>C00025+C20351	2.6.1.94	2.6.1A1(rev)|
R10096	C00029+C02627<=>C00015+C20353	2.4.1.284	2.4.1A1|
R10101	C00001+C00007+C15987<=>C00027+C00218+C00232	1.5.3.21	1.4.3A1|
R10102	C00001+C00007+C15987<=>C00027+C00067+C00334	1.5.3.19	1.4.3A1|
R10103	C00001+C00007+C20361<=>C00027+C00218+C19567	1.4.3.24	1.4.3A1|
R10105	C00001+C00006+C19567<=>C00005+C00080+C19569	1.2.1.83	1.2.1A2|
R10110	C00007+C19789<=>C20367	1.13.11.64	1.13.11A2|
R10114	C00006+C12448<=>C00005+C00080+C20371	1.1.1.334	1.1.1A2|1.2.1B2|
R10116	C00005+C00080+C20372<=>C00006+C20373	1.1.1.100	1.1.1A2(rev)|1.2.1B2(rev)|
R10117	C20373<=>C00001+C20374	4.2.1.59	4.2.1A1|4.2.1A2|4.2.1A6|4.2.1C1(rev)|
R10118	C00005+C00080+C20374<=>C00006+C20375	1.3.1.10	1.3.1A2(rev)|
R10120	C00005+C00080+C20376<=>C00006+C20377	1.1.1.100	1.1.1A2(rev)|1.2.1B2(rev)|
R10121	C20377<=>C00001+C20378	4.2.1.59	4.2.1A1|4.2.1A2|4.2.1A6|4.2.1C1(rev)|
R10122	C00005+C00080+C20378<=>C00006+C19846	1.3.1.10	1.3.1A2(rev)|
R10125	C00003+C09849<=>C00004+C00080+C09848	1.1.1.0	1.1.1A1|1.2.1B1|
R10128	C00003+C09131<=>C00004+C00080+C20379	1.1.1.332	1.1.1A1|1.2.1B1|
R10130	C00006+C17580<=>C00005+C00080+C17581	1.1.1.329	1.1.1A2|1.2.1B2|
R10134	C00001+C20383<=>C00191+C20380	3.1.1.93	3.1.1A1|3.2.1A1|
R10153	C00025<=>C05574	5.4.3.9	5.4.3A1|
R10161	C00003+C00100<=>C00004+C00080+C00894	1.3.1.95	1.3.1A1|
R10172	C01209+C01944<=>C00010+C00011+C05753	2.3.1.207	2.3.1D2|
R10178	C00022+C00334<=>C00041+C00232	2.6.1.96	2.6.1A1(rev)|
R10179	C00048+C00334<=>C00037+C00232	2.6.1.96	2.6.1A1(rev)|
R10180	C00022+C00078<=>C00041+C00331	2.6.1.99	2.6.1A1(rev)|
R10181	C00005+C00007+C00080+C00331<=>C00001+C00006+C00011+C00954	1.14.13.168	1.14.13A3|

R10190	C00003+C03187<=>C00004+C00080+C00688	1.1.1.339|1.1.1.344	1.1.1A1|1.2.1B1|
R10208	C01271<=>C00001+C00693	4.2.1.59	4.2.1A1|4.2.1A2|4.2.1A6|
R10212	C00001+C00007+C14279<=>C00027+C01546		1.2.3B1|
R10221	C00003+C00345<=>C00004+C00011+C00080+C00199	1.1.1.343|1.1.1.351	1.1.1.A9|
R10222	C00003+C02489<=>C00004+C00080+C00161	1.1.1.345	1.1.1A1|1.2.1B1|
R10232	C00001+C17325<=>C00014+C00868	3.5.4.35	3.5.4A1|6.3.4A1(rev)|
R10235	C00001+C01344<=>C00009+C06196	3.6.1.64	3.6.1A1|
R10250	C00019+C10453<=>C00021+C10454	2.1.1.146	2.1.1A5|
R10251	C00019+C16930<=>C00021+C10452	2.1.1.146	2.1.1A5|
R10253	C00024+C00590<=>C00010+C20225	2.3.1.0	2.3.1D1|2.3.1D1(rev)|
R10308	C00003+C01500<=>C00004+C00080+C01499	1.1.1.347	1.1.1A1|1.2.1B1|
R10332	C00003+C01217<=>C00004+C00080+C05927	1.5.1.47	1.5.1B1|
R10343	C00033+C00091<=>C00024+C00042	2.8.3.18	2.8.3A1(rev)|
R10412	C00003+C01126<=>C00004+C00080+C03461	1.1.1.354	1.1.1A1|1.2.1B1|
R10448	C00019+C00805<=>C00021+C12305	2.1.1.274	2.1.1A5|
R10458	C00019+C00688<=>C00021+C11460		2.1.1A5|
R10459	C00004+C00080+C11460<=>C00003+C12481		1.1.1A1(rev)|1.2.1B1(rev)|
R10466	C00001+C05933<=>C00077+C07044	3.5.3.25	3.5.4A1|
R10472	C00019+C10343<=>C00021+C01448	2.1.1.283	2.1.1A5|
R10473	C00024+C00556<=>C00010+C15513	2.3.1.224	2.3.1D1|2.3.1D1(rev)|
R10474	C00024+C02394<=>C00010+C12299	2.3.1.224	2.3.1D1|2.3.1D1(rev)|
R10491	C00019+C14155<=>C00021+C14153	2.1.1.95	2.1.1A5|
R10492	C00019+C14156<=>C00021+C14154	2.1.1.95	2.1.1A5|
R10520	C00003+C00092<=>C00004+C00080+C01236	1.1.1.363	1.1.1A1|1.2.1B1|
R10523	C00019+C00135<=>C00021+C03298	2.1.1.44	2.1.1A5|
R10524	C00019+C03298<=>C00021+C04259	2.1.1.44	2.1.1A5|
R10528	C00006+C02489<=>C00005+C00080+C00161	1.1.1.272	1.1.1A2|1.2.1B2|
R10547	C00001+C12213<=>C00009+C12212	3.1.3.92	3.1.3A1|
R10555	C01151<=>C01182	5.3.1.29	5.3.1A3|
R10561	C00002+C01107<=>C00008+C00009+C00011+C20345	4.1.1.0	4.1.1C1|
R10600	C00024+C00490<=>C00033+C00531	2.8.3.0	2.8.3A1(rev)|
R10612	C00001+C04348<=>C00010+C00149	3.1.2.30	3.1.2A1|3.3.1A1|6.2.1A1(rev)|6.2.1B1(rev)|6.2.1C1(rev)|6.2.1D1(rev)|
R10615	C00001+C00006+C00577<=>C00005+C00080+C00258	1.2.1.89	1.1.1D2|1.2.1A2|
R10638	C00029+C00561<=>C00015+C00844		2.4.1A1|
R10639	C00029+C00844<=>C00015+C08325		2.4.1A1|
R10687	C00001+C12270<=>C00025+C01042	3.4.17.21	3.5.1A1|6.3.1A1(rev)|
R10696	C00001+C09821<=>C09825	3.7.1.21	4.2.1A1(rev)|4.2.1A2(rev)|4.2.1A6(rev)|
R10699	C00047+C01092<=>C01037+C04076	2.6.1.105	2.6.1A1(rev)|
R10701	C00001+C10434<=>C00493+C01197	3.1.1.0	3.1.1A1|
R10703	C00003+C01845<=>C00004+C00080+C00207	1.1.1.0	1.1.1A1|1.2.1B1|
R10705	C00001+C17530<=>C00033+C00132	3.1.1.0	3.1.1A1|
R10707	C00024+C01209<=>C00010+C00011+C05744	2.3.1.180	2.3.1D2|
R10709	C00019+C17570<=>C00021+C16695	2.1.1.295	2.1.1A5|

Table A3: Thermochemical data for the reactants of isodesmic reactions of Table 7.2 that are used for the calculation of thermodynamic properties for the heavy metabolites in the gas phase

Compound	**$\Delta_f H$ / kJ mol^{-1}**	**$\Delta_f G$ / kJ mol^{-1}**
Methane	-74.52	-50.49
Ethane	-83.85	-31.92
Propane	-104.68	-24.39
n-Butane	-125.79	-16.7
isoButane	-134.18	-20.76
n-Pentane	-146.76	-8.813
isoPentane	-153.6	-14.05
Methanol	-200.94	-162.32
Ethanol	-234.95	-167.85
1-Propanol	-255.2	-159.9
2-Propanol	-272.7	-173.47
1-Butanol	-274.6	-150.3
Ethylene glycol	-387.5	-302.6
p-Ethylphenol	-144.05	-21.58
Acetaldehyde	-166.2	-132.8
Acetone	-215.7	-151.3
Acetic acid	-432.25	-374.6
Propanoic acid	-453.5	-366.7
Butanoic acid	-475.8	-360.0
Dimethyl ether	-183.94	-112.8
Methylamine	-19.38	32.07
Dimethylamine	-18.45	68.39
Propylamine	-70.1	41.7
Butylamine	-91.9	49.3
Pentylamine	-113.2	57.3
Ammonia	-45.57	-16.4
Benzene	82.88	129.6
Ethylbenzene	29.92	130.73
Tetrahydrofuran (THF)	-182.5	-79.69
Tetrahydropyran (THP)	-224.28	-80.37
Acetamide	-238.3	-159.53
Propanamide	-259.0	-151.05
Butanamide	-282.0	-132.13
Imidazole	131.5	192.75
Indole	156.6	237.3
Pyrrolidine	-3.6	114.7
Guanidine	27.95	95.94
Methyl propyl sulfide	-82.3	17.93
Adenine	225.7	348.67
Guanine	16	184.49
Cytosine	-69.5	50.24
Uracil	-301.5	-198.35
Thymine	-338	-192.27

Table A4: Thermochemical quantities of neutral, deprotonated, and protonated aminoacids. Literature values are shown in parenthesis. All quantities are given in $kJmol^{-1}$.

Metabolite	$\Delta_f G^0$			$\Delta_f H^0$		
	neutral	**deprotonated**	**protonated**	**neutral**	**deprotonated**	**protonated**
Alanine	-298.0	-411	347.0	-421.9 (-423.1)	-524 (-528.6)	209.0 (209.5)
Arginine	-129.9	-273	389.8	-398.9	-532	99.7
Asparagine	-411.4	-572	196.6	-612.0	-760	-21.0
Aspartic acid	-621.8	-830	9.23	-790.2	-991	-177.2
Cysteine	-272.8	-411	370.2	-396.5	-525	233.5
Glutamic acid	-631.0	-812	-26.0	-819.0	-997	-233.0
Glutamine	-395.8	-558	182.2	-626.0	-777	-68.0
Glycine	-300.1	-409	358.9	-391.7 (-391.6)	-491 (-494.1)	254.3 (256.0)
Histidine	-87.6	-252	475.5	-289.9 (-289.5)[a]	-447 (-451.3)[a]	264.1 (265.1)[a]
Isoleucine	-279.7	-394	348.3	-493.5 (-492.1)	-601 (-606.3)	120.5 (123.9)
Leucine	-280.6	-395	349.4	-495.2 (-494.8)	-602 (-606.8)	121.8 (125.0)
Lysine	-205.9	-329	355.1	-458.8	-576	80.2
Methionine	-248.0	-356	366.2	-427.8	-534	167.5
Phenylalanine	-142.2	-268	478.9	-323.5	-433	279.5
Proline	-221.8	-341	383.2	-391.1 (-385.9)	-499 (-497.3)	200.0 (207.4)
Serine	-433.5	-583	201.5	-581.5	-722	39.5
Threonine	-439.5	-592	187.5	-620.0	-763	-6.0
Tryptophan	-29.8	-155	574.2	-246.9	-358	341.1
Tyrosine	-288.4	-415	329.6	-489.0	-602	111.0
Valine	-282.2	-398	349.8	-472.6 (-473.4)	-577 (-586.7)	145.4 (146.7)

Table A5: Parameters of linear fits in Figures 7.5 and 7.6

Protonated Aminoacids				
	Free energy of formation (Fig. 7.5)		Enthalpy of formation (Fig. 7.6)	
R^2	0.953		0.943	
Intercept	589.0	SD = 15.8	583.7	SD = 27.9
Slope	0.925	SD = 0.047	0.976	SD = 0.055
Deprotonated Aminoacids				
R^2	0.984		0.986	
Intercept	-105.3	SD = 10.8	-64.5	SD = 15.7
Slope	1.107	SD = 0.032	1.131	SD = 0.031

Table A6: Parameters of linear fits in Figures 7.7 and 7.8

Protonated Aminoacids and References				
	$\Delta_f G^0$ (Line C, Fig. 7.7)		$\Delta_f H^0$ (Line F, Fig. 7.8)	
R^2	0.983		0.983	
Intercept	605.4	SD = 6.1	594.6	SD = 9.48
Slope	0.971	SD = 0.022	0.999	SD = 0.023
Deprotonated Aminoacids and References				
	$\Delta_f G^0$ (Line B, Fig. 7.7)		$\Delta_f H^0$ (Line E, Fig. 7.8)	
R^2	0.988		0.989	
Intercept	-113.5	SD = 7.0	--77.8	SD = 11.3
Slope	1.067	SD = 0.022	1.098	SD = 0.024
Protonated Sugars and References				
	$\Delta_f G^0$ (Line I, Fig. 7.7)		$\Delta_f H^0$ (Line L, Fig. 7.8)	
R^2	0.993		0.996	
Intercept	759.0	SD = 13.7	750.9	SD = 17.8
Slope	1.031	SD = 0.038	1.021	SD = 0.033
Deprotonated Sugars and References				
	$\Delta_f G^0$ (Line H, Fig. 7.7)		$\Delta_f H^0$ (Line K, Fig. 7.8)	
R^2	0.988		0.992	
Intercept	62.4	SD = 25.7	65.9	SD = 27.3
Slope	1.160	SD = 0.042	1.117	SD = 0.034

BIBLIOGRAPHY

1. United Nations. Department of International Economic and Social Affairs., United Nations. Department for Economic and Social Information and Policy Analysis., United Nations. Department of Economic and Social Affairs. Population Division.: **World urbanization prospects**. Edited by. New York: United Nations:volumes.
2. Wackett LP: **Biomass to fuels via microbial transformations**. *Current Opinion in Chemical Biology* 2008, **12**:187-193.
3. Kondo A, Ishii J, Hara KY, Hasunuma T, Matsuda F: **Development of microbial cell factories for bio-refinery through synthetic bioengineering**. *Journal of Biotechnology* 2013, **163**:204-216.
4. Colin VL, Rodriguez A, Cristobal HA: **The Role of Synthetic Biology in the Design of Microbial Cell Factories for Biofuel Production**. *Journal of Biomedicine and Biotechnology* 2011.
5. Caetano-Anolles G, Yafremava LS, Gee H, Caetano-Anolles D, Kim HS, Mittenthal JE: **The origin and evolution of modern metabolism**. *Int J Biochem Cell Biol* 2009, **41**:285-297.
6. Soh KC, Hatzimanikatis V: **DREAMS of metabolism**. *Trends in biotechnology* 2010, **28**:501-508.
7. Caspi R, Foerster H, Fulcher CA, Hopkinson R, Ingraham J, Kaipa P, Krummenacker M, Paley S, Pick J, Rhee SY, et al.: **MetaCyc: a multiorganism database of metabolic pathways and enzymes**. *Nucleic Acids Research* 2006, **34**:D511-D516.
8. Smith E, Morowitz HJ: **Universality in intermediary metabolism**. *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**:13168-13173.
9. Ehrenfreund P, Rasmussen S, Cleaves J, Chen L: **Experimentally tracing the key steps in the origin of life: The aromatic world**. *Astrobiology* 2006, **6**:490-520.
10. Tomar N, De RK: **Comparing methods for metabolic network analysis and an application to metabolic engineering**. *Gene* 2013, **521**:1-14.
11. Christensen B, Nielsen J: **Metabolic network analysis. A powerful tool in metabolic engineering**. *Adv Biochem Eng Biotechnol* 2000, **66**:209-231.
12. Murphy SA, Nicolaou A: **Lipidomics applications in health, disease and nutrition research**. *Molecular nutrition & food research* 2013, **57**:1336-1346.
13. Albinet V, Bats ML, Bedia C, Sabourdy F, Garcia V, Segui B, Andrieu-Abadie N, Hornemann T, Levade T: **Genetic disorders of simple sphingolipid metabolism**. *Handbook of experimental pharmacology* 2013:127-152.
14. Pralhada Rao R, Vaidyanathan N, Rengasamy M, Mammen Oommen A, Somaiya N, Jagannath MR: **Sphingolipid metabolic pathway: an overview of major roles played in human diseases**. *Journal of lipids* 2013, **2013**:178910.
15. Wenk MR: **The emerging field of lipidomics**. *Nature Reviews Drug Discovery* 2005, **4**:594-610.
16. Watson AD: **Thematic review series: systems biology approaches to metabolic and cardiovascular disorders. Lipidomics: a global approach to lipid analysis in biological systems**. *Journal of Lipid Research* 2006, **47**:2101-2111.
17. Hermansson M, Uphoff A, Kakela R, Somerharju P: **Automated quantitative analysis of complex lipidomes by liquid chromatography/mass spectrometry**. *Analytical Chemistry* 2005, **77**:2166-2175.
18. Lagarde M, Geloen A, Record M, Vance D, Spener F: **Lipidomics is emerging**. *Biochimica Et Biophysica Acta-Molecular and Cell Biology of Lipids* 2003, **1634**:61-61.

19. Zehethofer N, Pinto DM: **Recent developments in tandem mass spectrometry for lipidomic analysis**. *Analytica Chimica Acta* 2008, **627**:62-70.
20. Sokol E, Almeida R, Hannibal-Bach HK, Kotowska D, Vogt J, Baumgart J, Kristiansen K, Nitsch R, Knudsen J, Ejsing CS: **Profiling of lipid species by normal-phase liquid chromatography, nanoelectrospray ionization and ion trap-orbitrap mass spectrometry**. *Analytical biochemistry* 2013.
21. Yetukuri L, Katajamaa M, Medina-Gomez G, Seppanen-Laakso T, Vidal-Puig A, Oresic M: **Bioinformatics strategies for lipidomics analysis: characterization of obesity related hepatic steatosis**. *Bmc Systems Biology* 2007, **1**:12.
22. Oresic M: **Bioinformatics and computational approaches applicable to lipidomics**. *European Journal of Lipid Science and Technology* 2009, **111**:99-106.
23. Khalil MB, Hou WM, Zhou H, Elisma F, Swayne LA, Blanchard AP, Yao ZM, Bennett SAL, Figeys D: **Lipidomics Era: Accomplishments and Challenges**. *Mass Spectrometry Reviews* 2010, **29**:877-929.
24. Mann M, Nahar F, Ekker H, Backofen R, Stadler PF, Flamm C: **Atom Mapping with Constraint Programming**. *Principles and Practice of Constraint Programming, Cp 2013* 2013, **8124**:805-822.
25. Apostolakis J, Sacher O, Korner R, Gasteiger J: **Automatic determination of reaction mappings anal reaction center information. 2. Validation on a biochemical reaction database**. *Journal of Chemical Information and Modeling* 2008, **48**:1190-1198.
26. Mu F, Williams RF, Unkefer CJ, Unkefer PJ, Faeder JR, Hlavacek WS: **Carbon-fate maps for metabolic reactions**. *Bioinformatics* 2007, **23**:3193-3199.
27. Arita M: **In silico atomic tracing by substrate-product relationships in Escherichia coli intermediary metabolism**. *Genome Res* 2003, **13**:2455-2466.
28. Heath AP, Bennett GN, Kavraki LE: **Finding metabolic pathways using atom tracking**. *Bioinformatics* 2010, **26**:1548-1555.
29. Menkuc BS, Gille C, Holzhutter HG: **Computer aided optimization of carbon atom labeling for tracer experiments**. *Genome Inform* 2008, **20**:270-276.
30. Akutsu T: **Efficient extraction of mapping rules of atoms from enzymatic reaction data**. *Journal of Computational Biology* 2004, **11**:449-462.
31. Toya Y, Shimizu H: **Flux analysis and metabolomics for systematic metabolic engineering of microorganisms**. *Biotechnology Advances* 2013, **31**:818-826.
32. Sauer U: **Metabolic networks in motion: C-13-based flux analysis**. *Molecular Systems Biology* 2006, **2**.
33. Toya Y, Hirasawa T, Morimoto T, Masuda K, Kageyama Y, Ozaki K, Ogasawara N, Shimizu H: **C-13-metabolic flux analysis in heterologous cellulase production by Bacillus subtilis genome-reduced strain**. *Journal of Biotechnology* 2014, **179**:42-49.
34. Zamboni N, Fendt SM, Ruhl M, Sauer U: **C-13-based metabolic flux analysis**. *Nature Protocols* 2009, **4**:878-892.
35. Soh KC, Hatzimanikatis V: **Network thermodynamics in the post-genomic era**. *Curr Opin Microbiol* 2010, **13**:350-357.
36. Kasperski A: **Modelling of cells bioenergetics**. *Acta Biotheor* 2008, **56**:233-247.
37. Jankowski MD, Henry, C.S., Broadbelt, L.J., and Hatzimanikatis, V.: **Group contribution method for thermodynamic analysis of complex metabolic networks**. *Biophys J* 2008, **95**:1487-1499.
38. Price ND, J. A. Papin, C. H. Schilling, and B. O. Palsson: **Genome-scale microbial in silico models: the constraints-based approach.** *Trends Biotechnol.* 2003, **21**:162–169.
39. Henry CS, M. D. Jankowski, L. J. Broadbelt, and V. Hatzimanikatis.: **Genome-scale thermodynamic analysis of Escherichia coli metabolism**. *Biophys. J.* 2006, **90**:1453-1461.
40. Feist AM, C. S. Henry, J. L. Reed, M. Krummenacker, A. R. Joyce,, P. D. Karp LJB, V. Hatzimanikatis, and B. Ø. Palsson: **A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1261 ORFs and thermodynamic information**. *Mol. Syst. Biol.* 2007, **3**:1–18.

41. Kummel A, S. Panke, and M. Heinemann: **Systematic assignment of thermodynamic constraints in metabolic network models**. *BMC Bioinformatics* 2006, **7**:1–12.
42. Geyer T: **Modeling metabolic processes between molecular and systems biology**. *Current Opinion in Structural Biology* 2013, **23**:218-223.
43. Goldberg RN, Y. B. Tewari, and T. N. Bhat: **Thermodynamics of enzyme-catalyzed reactions—a database for quantitative biochemistry**. *Bioinformatics* 2004, **20**:2874–2877.
44. 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**. *Sci Rep* 2014, **4**:7022.
45. Beard DA, and H. Qian.: **Thermodynamic-based computational profiling of cellular regulatory control in hepatocyte metabolism**. *Am. J. Physiol. Endocrinol. Metab.* 2005, **288**:E633–E644.
46. Kummel A, S. Panke, and M. Heinemann: **Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data**. *Mol. Syst. Biol.* 2006, **2**:1–10.
47. Henry CS, Broadbelt LJ, Hatzimanikatis V: **Thermodynamics-based metabolic flux analysis**. *Biophysical journal* 2007, **92**:1792-1805.
48. Jankowski MD, Henry CS, Broadbelt LJ, Hatzimanikatis V: **Group contribution method for thermodynamic analysis of complex metabolic networks**. *Biophys J* 2008, **95**:1487-1499.
49. Jacquel N, Lo C-W, Wei Y-H, Wu H-S, Wang SS: **Isolation and purification of bacterial poly(3-hydroxyalkanoates)**. *Biochemical Engineering Journal* 2008, **39**:15-27.
50. B P: **Introduction to Systems Biology**. *Short Course on Systems Biology, Iceland* 2008:19.
51. Zhao XM, Tian WD, Jiang R, Wan J: **Computational Systems Biology**. *Scientific World Journal* 2013.
52. Medina MA: **Systems biology for molecular life sciences and its impact in biomedicine (vol 70, pg 1035, 2013)**. *Cellular and Molecular Life Sciences* 2013, **70**:3475-3480.
53. Stephanopoulos G: **Synthetic Biology and Metabolic Engineering**. *Acs Synthetic Biology* 2012, **1**:514-525.
54. Kobayashi H, Kaern M, Araki M, Chung K, Gardner TS, Cantor CR, Collins JJ: **Programmable cells: Interfacing natural and engineered gene networks**. *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**:8414-8419.
55. Shin JH, Kim HU, Kim DI, Lee SY: **Production of bulk chemicals via novel metabolic pathways in microorganisms**. *Biotechnology Advances* 2013, **31**:925-935.
56. Curran KA, Alper HS: **Expanding the chemical palate of cells by combining systems biology and metabolic engineering**. *Metabolic Engineering* 2012, **14**:289-297.
57. Ferrer P: **Systems biology and biological systems diversity for the engineering of microbial cell factories**. *Microb Cell Fact* 2007, **6**:35.
58. Yip YL, Famiglietti M, Gos A, Duek PD, David FP, Gateau A, Bairoch A: **Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase**. *Hum Mutat* 2008, **29**:361-366.
59. McDonald AG, Boyce S, Moss GP, Dixon HB, Tipton KF: **ExplorEnz: a MySQL database of the IUBMB enzyme nomenclature**. *BMC biochemistry* 2007, **8**:14.
60. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes**. *Nucleic acids research* 2000, **28**:27-30.
61. Karp PD, Riley M, Saier M, Paulsen IT, Paley SM, Pellegrini-Toole A: **The EcoCyc and MetaCyc databases**. *Nucleic Acids Research* 2000, **28**:56-59.
62. Zhou T: **Computational reconstruction of metabolic networks from KEGG**. *Methods Mol Biol* 2013, **930**:235-249.
63. 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.

64. Thiele I, Palsson BO: **A protocol for generating a high-quality genome-scale metabolic reconstruction**. *Nat Protoc* 2010, **5**:93-121.
65. Agren R, Liu L, Shoaie S, Vongsangnak W, Nookaew I, Nielsen J: **The RAVEN Toolbox and Its Use for Generating a Genome-scale Metabolic Model for <italic>Penicillium chrysogenum</italic>**. *PLoS Comput Biol* 2013, **9**:e1002980.
66. Schellenberger J, Park JO, Conrad TM, Palsson BO: **BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions**. *BMC Bioinformatics* 2010, **11**:-.
67. 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**.
68. King ZA, Lloyd CJ, Feist AM, Palsson BO: **Next-generation genome-scale models for metabolic engineering**. *Curr Opin Biotechnol* 2015, **35C**:23-29.
69. Chou CH, Chang WC, Chiu CM, Huang CC, Huang HD: **FMM: a web server for metabolic pathway reconstruction and comparative analysis**. *Nucleic Acids Research* 2009, **37**:W129-W134.
70. Faust K, Dupont P, Callut J, van Helden J: **Pathway discovery in metabolic networks by subgraph extraction**. *Bioinformatics* 2010.
71. Dale JM, Popescu L, Karp PD: **Machine learning methods for metabolic pathway prediction**. *BMC Bioinformatics* 2010, **11**:15.
72. Hatzimanikatis V, Li CH, Ionita JA, Henry CS, Jankowski MD, Broadbelt LJ: **Exploring the diversity of complex metabolic networks**. *Bioinformatics* 2005, **21**:1603-1609.
73. Cho A, Yun H, Park JH, Lee SY, Park S: **Prediction of novel synthetic pathways for the production of desired chemicals**. *Bmc Systems Biology* 2010, **4**.
74. Carbonell P, Planson AG, Fichera D, Faulon JL: **A retrosynthetic biology approach to metabolic pathway design for therapeutic production**. *Bmc Systems Biology* 2011, **5**.
75. Rodrigo G, Carrera J, Prather KJ, Jaramillo A: **DESHARKY: automatic design of metabolic pathways for optimal cell growth**. *Bioinformatics* 2008, **24**:2554-2556.
76. 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**. *Nat Chem Biol* 2011, **7**:445-452.
77. 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.
78. Prather KL, Martin CH: **De novo biosynthetic pathways: rational design of microbial chemical factories**. *Curr Opin Biotechnol* 2008, **19**:468-474.
79. Ellis LB, Roe D, Wackett LP: **The University of Minnesota Biocatalysis/Biodegradation Database: the first decade**. *Nucleic Acids Res* 2006, **34**:D517-521.
80. Pharkya P, Burgard AP, Maranas CD: **OptStrain: A computational framework for redesign of microbial production systems**. *Genome Research* 2004, **14**:2367-2376.
81. Hatzimanikatis V, Li CH, Ionita JA, Broadbelt LJ: **Metabolic networks: enzyme function and metabolite structure**. *Curr Opin in Struct Biol* 2004, **14**:300-306.
82. 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.
83. Brunk E, Neri M, Tavernelli I, Hatzimanikatis V, Rothlisberger U: **Integrating computational methods to retrofit enzymes to synthetic pathways**. *Biotechnol Bioeng* 2012, **109**:572-582.
84. 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.
85. Martin CH, Nielsen DR, Solomon KV, Prather KL: **Synthetic metabolism: engineering biology at the protein and pathway scales**. *Chem Biol* 2009, **16**:277-286.

86. Feher T, Planson AG, Carbonell P, Fernandez-Castane A, Grigoras I, Dariy E, Perret A, Faulon JL: **Validation of RetroPath, a computer-aided design tool for metabolic pathway engineering**. *Biotechnology Journal* 2014, **9**:1446-1457.
87. 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.
88. Planson AG, Carbonell P, Paillard E, Pollet N, Faulon JL: **Compound toxicity screening and structure-activity relationship modeling in Escherichia coli**. *Biotechnol Bioeng* 2012, **109**:846-850.
89. Carbonell P, Fichera D, Pandit SB, Faulon JL: **Enumerating metabolic pathways for the production of heterologous target chemicals in chassis organisms**. *Bmc Systems Biology* 2012, **6**.
90. Carbonell P, Planson AG, Faulon JL: **Retrosynthetic design of heterologous pathways**. *Methods Mol Biol* 2013, **985**:149-173.
91. Metzenberg RL: **Norman Harold Horowitz, 1915-2005**. *Genetics* 2005, **171**:1445-1448.
92. Bachmann BO: **Biosynthesis: is it time to go retro?** *Nat Chem Biol* 2010, **6**:390-393.
93. Corey EJ: **The Logic of Chemical Synthesis - Multistep Synthesis of Complex Carbogenic Molecules**. *Angewandte Chemie-International Edition in English* 1991, **30**:455-465.
94. Law J, Zsoldos Z, Simon A, Reid D, Liu Y, Khew SY, Johnson AP, Major S, Wade RA, Ando HY: **Route Designer: A Retrosynthetic Analysis Tool Utilizing Automated Retrosynthetic Rule Generation**. *Journal of Chemical Information and Modeling* 2009, **49**:593-602.
95. Finley SD, Broadbelt LJ, Hatzimanikatis V: **Computational Framework for Predictive Biodegradation**. *Biotechnology and Bioengineering* 2009, **104**:1086-1097.
96. Planson AG, Carbonell P, Grigoras I, Faulon JL: **A retrosynthetic biology approach to therapeutics: from conception to delivery**. *Current Opinion in Biotechnology* 2012, **23**:948-956.
97. Lee JW, Na D, Park JM, Lee J, Choi S, Lee SY: **Systems metabolic engineering of microorganisms for natural and non-natural chemicals**. *Nature Chemical Biology* 2012, **8**:536-546.
98. Martin CH, Nielsen DR, Solomon KV, Prather KLJ: **Synthetic Metabolism: Engineering Biology at the Protein and Pathway Scales**. *Chemistry & Biology* 2009, **16**:277-286.
99. Long MR, Ong WK, Reed JL: **Computational methods in metabolic engineering for strain design**. *Curr Opin Biotechnol* 2015, **34C**:135-141.
100. Barrett AJ: **Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Enzyme Nomenclature. Recommendations 1992. Supplement 4: corrections and additions (1997)**. *European journal of biochemistry / FEBS* 1997, **250**:1-6.
101. Ugi I, Bauer J, Brandt J, Friedrich J, Gasteiger J, Jochum C, Schubert W: **New Applications of Computers in Chemistry**. *Angewandte Chemie-International Edition in English* 1979, **18**:111-123.
102. Gonzalez-Lergier J, Broadbelt LJ, Hatzimanikatis V: **Theoretical considerations and computational analysis of the complexity in polyketide synthesis pathways**. *Journal of the American Chemical Society* 2005, **127**:9930-9938.
103. Hatzimanikatis V, Li C, Ionita JA, Broadbelt LJ: **Metabolic networks: enzyme function and metabolite structure**. *Curr Opin Struct Biol* 2004, **14**:300-306.
104. 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.
105. Degtyarenko K, Hastings J, de Matos P, Ennis M: **ChEBI: an open bioinformatics and cheminformatics resource**. *Curr Protoc Bioinformatics* 2009, **Chapter 14**:Unit 14 19.
106. Faust K, Croes D, van Helden J: **Metabolic Pathfinding Using RPAIR Annotation**. *Journal of Molecular Biology* 2009, **388**:390-414.

107. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH: **PubChem: a public information system for analyzing bioactivities of small molecules**. *Nucleic Acids Res* 2009, **37**:W623-633.
108. Han L, Wang Y, Bryant SH: **Developing and validating predictive decision tree models from mining chemical structural fingerprints and high-throughput screening data in PubChem**. *BMC Bioinformatics* 2008, **9**:401.
109. Mavrovouniotis M, Stephanopoulos G, Stephanopoulos G: **Synthesis of Biochemical Production Routes**. *Computers & Chemical Engineering* 1992, **16**:605-619.
110. Yousofshahi M, Lee K, Hassoun S: **Probabilistic pathway construction**. *Metab Eng* 2011, **13**:435-444.
111. Engel T: **The structural- and bioassay database PubChem**. *Nachrichten Aus Der Chemie* 2007, **55**:521-524.
112. Degtyarenko K, De Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, Alcantara R, Darsow M, Guedj M, Ashburner M: **ChEBI: a database and ontology for chemical entities of biological interest**. *Nucleic Acids Research* 2008, **36**:D344-D350.
113. 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**.
114. Jankowski MD, Henry CS, Broadbelt LJ, Hatzimanikatis V: **Group contribution method for thermodynamic analysis of complex metabolic networks**. *Biophysical Journal* 2008, **95**:1487-1499.
115. Finley SD, Broadbelt LJ, Hatzimanikatis V: **In silico feasibility of novel biodegradation pathways for 1,2,4-trichlorobenzene**. *Bmc Systems Biology* 2010, **4**.
116. Kumar VS, Maranas CD: **GrowMatch: An Automated Method for Reconciling In Silico/In Vivo Growth Predictions**. *PLoS Comp Biol* 2009, **5**.
117. Gonzalez-Lergier J, Broadbelt LJ, Hatzimanikatis V: **Analysis of the maximum theoretical yield for the synthesis of erythromycin precursors in Escherichia coli**. *Biotechnol Bioeng* 2006, **95**:638-644.
118. Godden J, Xue L, Bajorath J: **Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and Tanimoto coefficients**. *J Chem Inf Comput Sci* 2000, **40**:163 - 166.
119. Chen, Vitkup: **Predicting genes for orphan metabolic activities using phylogenetic profiles**. *Genome Biol* 2006, **7**:R17.
120. Chen WL, Chen DZ, Taylor KT: **Automatic reaction mapping and reaction center detection**. *Wiley Interdisciplinary Reviews-Computational Molecular Science* 2013, **3**:560-593.
121. Lynch MF, Willett P: **Automatic Detection of Chemical-Reaction Sites**. *Journal of Chemical Information and Computer Sciences* 1978, **18**:154-159.
122. Korner R, Apostolakis J: **Automatic determination of reaction mappings and reaction center information. 1. The imaginary transition state energy approach**. *Journal of Chemical Information and Modeling* 2008, **48**:1181-1189.
123. Arita M: **In silico atomic tracing by substrate-product relationships in Escherichia coli intermediary metabolism**. *Genome Research* 2003, **13**:2455-2466.
124. Raymond JW, Willett P: **Maximum common subgraph isomorphism algorithms for the matching of chemical structures**. *Journal of Computer-Aided Molecular Design* 2002, **16**:521-533.
125. Kumar A, Maranas CD: **CLCA: Maximum Common Molecular Substructure Queries within the MetRxn Database**. *Journal of Chemical Information and Modeling* 2014, **54**:3417-3438.
126. Ravikirthi P, Suthers PF, Maranas CD: **Construction of an E. Coli genome-scale atom mapping model for MFA calculations**. *Biotechnol Bioeng* 2011, **108**:1372-1382.
127. Ugi IK, Wochner M: **Applications of the Principle of Minimal Chemical Distance**. *Abstracts of Papers of the American Chemical Society* 1988, **196**:47-COMP.

128. Crabtree JD, Mehta DP, Kouri TM: **An Open-Source Java Platform for Automated Reaction Mapping**. *Journal of Chemical Information and Modeling* 2010, **50**:1751-1756.
129. Heinonen M, Lappalainen S, Mielikainen T, Rousu J: **Computing Atom Mappings for Biochemical Reactions without Subgraph Isomorphism**. *Journal of Computational Biology* 2011, **18**:43-58.
130. First EL, Gounaris CE, Floudas CA: **Stereochemically consistent reaction mapping and identification of multiple reaction mechanisms through integer linear optimization**. *J Chem Inf Model* 2012, **52**:84-92.
131. Latendresse M, Malerich JP, Travers M, Karp PD: **Accurate Atom-Mapping Computation for Biochemical Reactions**. *J Chem Inf Model* 2012.
132. Muller C, Marcou G, Horvath D, Aires-de-Sousa J, Varnek A: **Models for identification of erroneous atom-to-atom mapping of reactions performed by automated algorithms**. *Journal of chemical information and modeling* 2012, **52**:3116-3122.
133. Latendresse M, Krummenacker M, Karp PD: **Optimal metabolic route search based on atom mappings**. *Bioinformatics* 2014, **30**:2043-2050.
134. Fooshee D, Andronico A, Baldi P: **ReactionMap: an efficient atom-mapping algorithm for chemical reactions**. *Journal of chemical information and modeling* 2013, **53**:2812-2819.
135. Pey J, Planes FJ, Beasley JE: **Refining carbon flux paths using atomic trace data**. *Bioinformatics* 2013.
136. Kotera M, Okuno Y, Hattori M, Goto S, Kanehisa M: **Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions**. *Journal of the American Chemical Society* 2004, **126**:16487-16498.
137. Mann M, Nahar F, Schnorr N, Backofen R, Stadler PF, Flamm C: **Atom mapping with constraint programming**. *Algorithms for Molecular Biology* 2014, **9**.
138. Mišković L, Hatzimanikatis V: **Modeling of uncertainties in biochemical reactions**. *Biotechnology and bioengineering* 2011, **108**:413-423.
139. Moreno-Sánchez R, Saavedra E, Rodríguez-Enríquez S, Olín-Sandoval V: **Metabolic control analysis: a tool for designing strategies to manipulate metabolic pathways**. *Journal of biomedicine & biotechnology* 2008, **2008**:597913-597913.
140. Mullins E, Liu YA, Ghaderi A, Fast SD: **Sigma Profile Database for Predicting Solid Solubility in Pure and Mixed Solvent Mixtures for Organic Pharmacological Compounds with COSMO-Based Thermodynamic Methods**. 2008:1707-1725.
141. Müller AC, Bockmayr A: **Flux modules in metabolic networks**. *Journal of mathematical biology* 2014, **69**:1151-1179.
142. van Speybroeck V, Gani R, Meier RJ: **The calculation of thermodynamic properties of molecules**. *Chemical Society Reviews* 2010, **39**:1764-1779.
143. Curtiss LAR, P. C.; Raghavachari, K.: **Gaussian-3X (G3X) theory: use of improved geometries, zero-point energies, and Hartree-Fock basis sets.** *J. Chem. Phys.* 2001, **114**:108–117.
144. Raghavachari KC, L. A. (Ed): *Quantum-Chemical Methods for Accurate Theoretical Thermochemistry* New York: Kluwer; 2002.
145. Curtiss LA, Redfern PC, Raghavachari K: **Assessment of Gaussian-4 theory for energy barriers**. *Chemical Physics Letters* 2010, **499**:168-172.
146. Dorofeeva OV, Ryzhova ON, Suntsova MA: **Accurate Prediction of Enthalpies of Formation of Organic Azides by Combining G4 Theory Calculations with an Isodesmic Reaction Scheme**. *Journal of Physical Chemistry A* 2013, **117**:6835-6845.
147. Ruscic B, Pinzon RE, Morton ML, von Laszevski G, Bittner SJ, Nijsure SG, Amin KA, Minkoff M, Wagner AF: **Introduction to active thermochemical tables: Several "key" enthalpies of formation revisited**. *Journal of Physical Chemistry A* 2004, **108**:9979-9997.
148. Elke Goos ABaBR: **Extended Third Millennium Ideal Gas and Condensed Phase Thermochemical Database for Combustion with updates from Active Thermochemical Tables**. 2014.

149. Thomson GH: **Experimental project results from the Design Institute for Physical Property Data (DIPPR) of the American Institute of Chemical Engineers. 4**. *Journal of Chemical and Engineering Data* 2000, **45**:145-145.
150. Gronert S, Ohair RAJ: **Ab-Initio Studies of Amino-Acid Conformations .1. The Conformers of Alanine, Serine, and Cysteine**. *Journal of the American Chemical Society* 1995, **117**:2071-2081.
151. Gronert S, Simpson DC, Conner KM: **A Reevaluation of Computed Proton Affinities for the Common alpha-Amino Acids**. *Journal of the American Society for Mass Spectrometry* 2009, **20**:2116-2123.
152. Hunter EP LS: **Evaluated gas phase basicities and proton affinities of molecules: An update.** *J Phys Chem Ref Data* 1998, **27**:413–656.
153. Harrison AG: **The gas-phase basicities and proton affinities of amino acids and peptides**. *Mass Spectrometry Reviews* 1997, **16**:201-217.
154. Hahn IS, Wesdemiotis C: **Protonation thermochemistry of beta-alanine - An evaluation of proton affinities and entropies determined by the extended kinetic method**. *International Journal of Mass Spectrometry* 2003, **222**:465-479.
155. Drahos L, Peltz C, Vekey K: **Accuracy of enthalpy and entropy determination using the kinetic method: are we approaching a consensus?** *Journal of Mass Spectrometry* 2004, **39**:1016-1024.
156. Bouchoux G: **Gas-phase basicities of polyfunctional molecules. part 1: Theory and methods**. *Mass Spectrometry Reviews* 2007, **26**:775-835.
157. Stover ML, Jackson VE, Matus MH, Adams MA, Cassady CJ, Dixon DA: **Fundamental Thermochemical Properties of Amino Acids: Gas-Phase and Aqueous Acidities and Gas-Phase Heats of Formation**. *Journal of Physical Chemistry B* 2012, **116**:2905-2916.
158. Bouchoux G: **Gas phase basicities of polyfunctional molecules. Part 3: Amino acids**. *Mass Spectrometry Reviews* 2012, **31**:391-435.
159. Jones CM, Bernier M, Carson E, Colyer KE, Metz R, Pawlow A, Wischow ED, Webb I, Andriole EJ, Poutsma JC: **Gas-phase acidities of the 20 protein amino acids**. *International Journal of Mass Spectrometry* 2007, **267**:54-62.
160. Rak J, Skurski P, Simons J, Gutowski M: **Low-energy tautomers and conformers of neutral and protonated arginine**. *Journal of the American Chemical Society* 2001, **123**:11695-11707.
161. Grimme S, Ehrlich S, Goerigk L: **Effect of the Damping Function in Dispersion Corrected Density Functional Theory**. *Journal of Computational Chemistry* 2011, **32**:1456-1465.
162. Von Arnim M, Ahlrichs R: **Performance of parallel TURBOMOLE for density functional calculations**. *Journal of Computational Chemistry* 1998, **19**:1746-1757.
163. Eichkorn K, Weigend F, Treutler O, Ahlrichs R: **Auxiliary basis sets for main row atoms and transition metals and their use to approximate Coulomb potentials**. *Theoretical Chemistry Accounts* 1997, **97**:119-124.
164. Lee CT, Yang WT, Parr RG: **Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron-Density**. *Physical Review B* 1988, **37**:785-789.
165. Becke AD: **Density-Functional Thermochemistry .3. The Role of Exact Exchange**. *Journal of Chemical Physics* 1993, **98**:5648-5652.
166. Becke AD: **Density-functional thermochemistry .5. Systematic optimization of exchange-correlation functionals**. *Journal of Chemical Physics* 1997, **107**:8554-8560.
167. Grimme S: **Semiempirical GGA-type density functional constructed with a long-range dispersion correction**. *Journal of Computational Chemistry* 2006, **27**:1787-1799.
168. Chai JD, Head-Gordon M: **Long-range corrected hybrid density functionals with damped atom-atom dispersion corrections**. *Physical Chemistry Chemical Physics* 2008, **10**:6615-6620.
169. Boerio-Goates JBOaJ: *Chemical Thermodynamics: Principles and Applications*; 2000.

170. Orth JD, Palsson B: **Gap-filling analysis of the iJO1366 Escherichia coli metabolic network reconstruction for discovery of metabolic functions**. *BMC Syst Biol* 2012, **6**:30.
171. Nohturfft A, Zhang SC: **Coordination of lipid metabolism in membrane biogenesis**. *Annual review of cell and developmental biology* 2009, **25**:539-566.
172. Pulfer M, Murphy RC: **Electrospray mass spectrometry of phospholipids**. *Mass Spectrometry Reviews* 2003, **22**:332-364.
173. Hanahan DJ: *A guide to phospholipid chemistry*. New York ; Oxford: Oxford University Press; 1997.
174. Fahy E, Subramaniam S, Brown HA, Glass CK, Merrill AH, Jr., Murphy RC, Raetz CR, Russell DW, Seyama Y, Shaw W, et al.: **A comprehensive classification system for lipids**. *Journal of Lipid Research* 2005, **46**:839-861.
175. Fahy E, Sud M, Cotter D, Subramaniam S: **LIPID MAPS online tools for lipid research**. *Nucleic acids research* 2007, **35**:W606-W612.
176. Fahy E, Cotter D, Byrnes R, Sud M, Maer A, Li J, Nadeau D, Zhau Y, Subramaniam S: **Bioinformatics for lipidomics**. *Methods in enzymology* 2007, **432**:247-273.
177. Goursot A, Mineva T, Bissig C, Gruenberg J, Salahub DR: **Structure, dynamics, and energetics of lysobisphosphatidic acid (LBPA) isomers**. *The journal of physical chemistry. B* 2010, **114**:15712-15720.
178. Hullin-Matsuda F, Kawasaki K, Delton-Vandenbroucke I, Xu Y, Nishijima M, Lagarde M, Schlame M, Kobayashi T: **De novo biosynthesis of the late endosome lipid, bis(monoacylglycero)phosphate**. *Journal of Lipid Research* 2007, **48**:1997-2008.
179. Ullmann JR: **Algorithm for Subgraph Isomorphism**. *Journal of the Acm* 1976, **23**:31-42.
180. El-Sonbaty Y, Ismail MA: **A new algorithm for subgraph optimal isomorphism**. *Pattern Recognition* 1998, **31**:205-218.
181. Eppstein D: **Subgraph Isomorphism in Planar Graphs and Related Problems**. *Proceedings of the Sixth Annual Acm-Siam Symposium on Discrete Algorithms* 1995:632-640
654.
182. Cordella LP, Foggia P, Sansone C, Vento M: **A (sub)graph isomorphism algorithm for matching large graphs**. *Ieee Transactions on Pattern Analysis and Machine Intelligence* 2004, **26**:1367-1372.
183. P. Foggia CS, M. Vento: **A Performance Comparison of Five Algorithms for Graph Isomorphism**. In *3rd IAPR TC-15 Workshop on Graph-based Representations in Pattern Recognition*. Edited by; 2001.
184. Wild DJ, Blankley CJ: **Comparison of 2D fingerprint types and hierarchy level selection methods for structural grouping using Ward's clustering**. *J Chem Inf Comput Sci* 2000, **40**:155-162.
185. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR: **Open Babel: An open chemical toolbox**. *J Cheminform* 2011, **3**:33.
186. Finley SD, Broadbelt LJ, Hatzimanikatis V: **In silico feasibility of novel biodegradation pathways for 1,2,4-trichlorobenzene**. *BMC systems biology* 2010, **4**:7.
187. Finley SD, Broadbelt LJ, Hatzimanikatis V: **Thermodynamic analysis of biodegradation pathways**. *Biotechnology and bioengineering* 2009, **103**:532-541.
188. Mavrovouniotis ML: **Group contributions for estimating standard gibbs energies of formation of biochemical compounds in aqueous solution**. *Biotechnology and bioengineering* 1990, **36**:1070-1082.
189. Hullin-Matsuda F, Luquain-Costaz C, Bouvier J, Delton-Vandenbroucke I: **Bis(monoacylglycero)phosphate, a peculiar phospholipid to control the fate of cholesterol: Implications in pathology**. *Prostaglandins Leukotrienes and Essential Fatty Acids* 2009, **81**:313-324.
190. Feist A, Henry C, Reed J, Krummenacker M, Joyce A, Karp P, Broadbelt L, Hatzimanikatis V, Palsson B: **A genome-scale metabolic reconstruction for Escherichia coli K-12**

MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol* 2007, **3**:121.
191. Hadadi N, Ataman M, Hatzimanikatis V, Panayiotou C: **Molecular thermodynamics of metabolism: quantum thermochemical calculations for key metabolites**. *Phys Chem Chem Phys* 2015.
192. 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.
193. Soh KC, Hatzimanikatis V: **Network thermodynamics in the post-genomic era**. *Current opinion in microbiology* 2010, **13**:350-357.
194. 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**:121.
195. Dalby A, Nourse JG, Hounshell WD, Gushurst AKI, Grier DL, Leland BA, Laufer J: **Description of Several Chemical-Structure File Formats Used by Computer-Programs Developed at Molecular Design Limited**. *Journal of Chemical Information and Computer Sciences* 1992, **32**:244-255.
196. Lilley DMJ, Clegg RM, Diekmann S, Seeman NC, Vonkitzing E, Hagerman P: **Nomenclature Committee of the International Union of Biochemistry and Molecular-Biology (Nc-Iubmb) - a Nomenclature of Junctions and Branchpoints in Nucleic-Acids - Recommendations 1994**. *European Journal of Biochemistry* 1995, **230**:1-2.
197. Cai YD, Huang T, Chen L, Niu B: **Application of Systems Biology and Bioinformatics Methods in Biochemistry and Biomedicine**. *Biomed Research International* 2013.
198. Cohen N: **Revised group additivity values for enthalpies of formation (at 298 K) of carbon-hydrogen and carbon-hydrogen-oxygen compounds**. *Journal of Physical and Chemical Reference Data* 1996, **25**:1411-1481.
199. Domalski ES, Hearing ED: **Estimation of the Thermodynamic Properties of C-H-N-O-S-Halogen Compounds at 298.15-K**. *Journal of Physical and Chemical Reference Data* 1993, **22**:805-1159.
200. Sagadeev EV, Gimadeev AA, Chachkov DV, Barabanov VP: **Empirical and ab initio Calculations of Thermochemical Parameters of Aminoacids: II. Diaminomonocarboxylic Acids, Hydroxyamino Acids, Thioamino Acids, and Heterocyclic Amino(imino) Acids**. *Russian Journal of General Chemistry* 2009, **79**:1490-1493.
201. Mavrovouniotis ML: **Estimation of standard Gibbs energy changes of biotransformations**. *J. Biol. Chem.* 1991, **266**:14440–14445.
202. Amend JP, Helgeson HC: **Group additivity equations of state for calculating the standard molal thermodynamic properties of aqueous organic species at elevated temperatures and pressures**. *Geochimica Et Cosmochimica Acta* 1997, **61**:11-46.
203. Plyasunov AV, Shock EL: **Thermodynamic functions of hydration of hydrocarbons at 298.15 K and 0.1 MPa**. *Geochimica Et Cosmochimica Acta* 2000, **64**:439-468.
204. Alberty RA: *Thermodynamics of Biochemical Reactions.* Cambridge, MA.: Massachusetts Institute of Technology Press; 2003.
205. Scholten JCM, J. C. Murrell, and D. P. Kelly: **Growth of sulfate-reducing bacteria and methanogenic archaea with methylated sulfur compounds: a commentary on the thermodynamic aspects**. *Arch. Microbiol* 2003, **179**:135–144.
206. Tanaka M, Okuno, Y., Yamada, T., Goto, S., Uemura, S. and Kanehisa, M.: **Extraction of a thermodynamic property for biochemical reactions in the metabolic pathway**. *Genome Inform* 2003, **14**:370–371.
207. Li C, J. A. Ionita, C. S. Henry, M. D. Jankowski, V. Hatzimanikatis,, Broadbelt aLJ: **Computational discovery of biochemical routes to specialty chemicals**. *Chem. Eng. Sci* 2004, **59**:5051–5060.

208. Tewari YB, Goldberg RN: **Thermodynamics of the hydrolysis reactions of nitriles**. *Journal of Chemical Thermodynamics* 2005, **37**:720-728.
209. LaRowe DEaH, H. C.: **Biomolecules in hydrothermal systems: Calculation of the standard molal thermodynamic properties of nucleic-acid bases, nucleosides, and nucleotides at elevated pressures and pressures**. *Geochimica Cosmochimica Acta* 2006, **70**:4680–4724.
210. Picard AaD, I.: **Pressure as an environmental parameter for microbial life — A review**. *Biophys. Chem.* 2013, **183**:30–41.
211. Sameera WMC, Pantazis DA: **A Hierarchy of Methods for the Energetically Accurate Modeling of Isomerism in Monosaccharides**. *Journal of Chemical Theory and Computation* 2012, **8**:2630-2645.
212. Missopolinou D, Panayiotou C: **Hydrogen-bonding cooperativity and competing inter- and intramolecular associations: A unified approach**. *Journal of Physical Chemistry A* 1998, **102**:3574-3581.
213. Riffet V, Bouchoux G: **Gas-phase structures and thermochemistry of neutral histidine and its conjugated acid and base**. *Physical Chemistry Chemical Physics* 2013, **15**:6097-6106.
214. Mayes HB, Broadbelt LJ, Beckham GT: **How Sugars Pucker: Electronic Structure Calculations Map the Kinetic Landscape of Five Biologically Paramount Monosaccharides and Their Implications for Enzymatic Catalysis**. *Journal of the American Chemical Society* 2014, **136**:1008-1022.
215. Linstrom PJ, Mallard WG: **The NIST Chemistry WebBook: A chemical data resource on the internet**. *Journal of Chemical and Engineering Data* 2001, **46**:1059-1063.
216. Trygubenko SAB, T. V.; Rueda, M.; Orozco, M.; Luque, F. J.; Sponer, J.; Slavicek, P.; Hobza, P.: **Correlated ab Initio Study of Nucleic Acid Bases and Their Tautomers in the Gas Phase, in a Microhydrated Environment and in Aqueous Solution. Part 1. Cytosine**. *Phys. Chem. Chem. Phys.* 2002, **4**:4192–4203.
217. Hanus MR, F.; Kabelac, M.; Kubar, T.; Bogdan, T. V.; Trygubenko, S. A.; Hobza, P.: **Correlated ab Initio Study of Nucleic Acid Bases and Their Tautomers in the Gas Phase, in a Microhydrated Environment and in Aqueous Solution. Guanine: Surprising Stabilization of Rare Tautomers in Aqueous Solution**. *J. Am. Chem. Soc.* 2003, **125**:7678–7688.
218. Alecu IM, Zheng JJ, Zhao Y, Truhlar DG: **Computational Thermochemistry: Scale Factor Databases and Scale Factors for Vibrational Frequencies Obtained from Electronic Model Chemistries**. *Journal of Chemical Theory and Computation* 2010, **6**:2872-2887.
219. Dorofeeva OV, Ryzhova ON: **Revision of standard molar enthalpies of formation of glycine and L-alanine in the gaseous phase on the basis of theoretical calculations**. *Journal of Chemical Thermodynamics* 2009, **41**:433-438.
220. Ngauv SN, Sabbah R, Laffitte M: **Thermodynamics of Nitrogen-Compounds .3. Thermochemical Study of Glycine and L-Alpha-Alanine**. *Thermochimica Acta* 1977, **20**:371-380.
221. Dorofeeva OV, Kolesnikova IN, Marochkin II, Ryzhova ON: **Assessment of Gaussian-4 theory for the computation of enthalpies of formation of large organic molecules**. *Structural Chemistry* 2011, **22**:1303-1314.
222. Pedley B (Ed): *Thermochemical Data and Structures of Organic Compounds* College Station, TX: TRC; 1994.
223. Sagadeev EV, Gimadeev AA, Barabanov VP: **The Enthalpies of Formation and Sublimation of Amino Acids and Peptides**. *Russian Journal of Physical Chemistry A* 2010, **84**:209-214.
224. Chase MWJ: *NIST-JANAF Thermochemical Tables, 4th ed.* Woodbury, NY: American Institute of Physics; 1998.
225. Sabbah R, Laffitte M: **Thermodynamics of Nitrogen-Compounds .4. Thermochemical Study of Sarcosine and L-Proline**. *Bulletin De La Societe Chimique De France Partie I-*

Physicochimie Des Systemes Liquides Electrochimie Catalyse Genie Chimique 1978:I50-I52.
226. Jang YS, Park JM, Choi S, Choi YJ, Seung DY, Cho JH, Lee SY: **Engineering of microorganisms for the production of biofuels and perspectives based on systems metabolic engineering approaches**. *Biotechnology Advances* 2012, **30**:989-1000.
227. Santos AF, Ribeiro da Silva MA: **Experimental and computational study on the molecular energetics of 2-pyrrolecarboxylic acid and 1-methyl-2-pyrrolecarboxylic acid**. *The journal of physical chemistry. A* 2009, **113**:9741-9750.
228. Jebber KA, Zhang K, Cassady CJ, ChungPhillips A: **Ab initio and experimental studies on the protonation of glucose in the gas phase**. *Journal of the American Chemical Society* 1996, **118**:10515-10524.

Noushin HADADI

Address: Chemin de Crissier 16, 1008 Jouxtens-Mézery (VD), Switzerland
Phone: +41 78 753 94 39
E-mail: noushin.hadadi@gmail.com
Personal information: 30 years old, married, Iranian, work permit B (CH)

- PhD in Systems Biotechnology, MSc in Biochemical Engineering
- Strong organizational and analytical skills
- Good communication, teamwork / project management and leadership

EDUCATION

July 2015 - Present
Postdoctoral Research Assistant
École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
Laboratory of Computational Systems Biotechnology (LCSB)

2010 – 2015
Ph.D in Systems Biotechnology
École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
Doctoral Program in Biotechnology and Bioengineering
Ph.D thesis: *Computational Studies on Cellular Metabolism: From Biochemical Pathways to Complex Metabolic Networks*
Advisor: Professor Vassily Hatzimanikatis

2008 – 2010
M.Sc. in Biochemical Engineering & Biotechnology
École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
Master thesis: *Exploring Novel Pathways and Structures in Lipid Metabolism Using Computational Methods*

2003 – 2008
B.Sc. in Chemical Engineering
University of Tehran, Iran

PROFESSIONAL EXPERIENCE

November 2014 – April 2015
Research Intern, Nestlé Institute of Health Sciences (NIHS)
- Statistical analysis of metagenomics data
- Study and analysis of the fundamental mechanisms underlying the host-microbes interactions in living organisms

July 2010 – Present
Research Assistant and Ph.D Candidate, **EPFL**
- Using the engineering fundamentals of process design and computational methods for *in silico* generation and analysis of novel metabolic pathways for metabolic engineering and other applications
- Developing computational methods to discover the full metabolic capacity of model microorganisms (namely *E. coli* and *yeast*)
- Computational studies on lipid metabolism and developing computational tools for providing guidance on designing novel therapeutic interventions for lipid-associated disorders
- 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” and “Principles and Applications of Systems Biology” courses and supervising five bachelor and master level students thesis and projects

July 2009 – January 2010
Intern
École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
Cellular Biotechnology Laboratory (LBTC)
Project: Worked on optimizing the parameters influencing the performance of bio-reactors in cell culture processes

COMPUTER SKILLS

Operating Systems: Mac OS, Windows, LINUX
Programming: MATLAB, C++, Python
Process simulators: HYSYS, ASPEN
Database: MySQL
Productivity Software: Microsoft Office (Word, Excel, Powerpoint), LATEX

LANGUAGE SKILLS

English: Proficient, **French:** having level B2 certificate, **Persian:** Native speaker

RELEVANT COURSES

Master courses, EPFL
Advanced biotechnology, Risk management, Chemical process safety, Equilibrium-stage separation processes, Process development, Biochemical engineering, Principles and applications of systems biology.

Doctoral courses, EPFL
Biomedical approaches for drug evaluation, Functional genomics, Advanced mass spectrometry, Scientific English Writing.

PUBLICATIONS

Hadadi, N., Soh K.C., Seijo M., Zisaki A., Guan X., Wenk M., Hatzimanikatis V. (2014) "A computational framework for integration of lipidomics data into metabolic pathways", *Metabolic engineering, vol. 23:1-8*
Hadadi, N., M. Ataman, V. Hatzimanikatis and C. Panayiotou**.** (2015) **"**Molecular thermodynamics of metabolism: quantum thermochemical calculations for key metablites", *Physical Chemistry Chemical Physics*, vol. 17: 10438-10453.
Hadadi, N. and Hatzimanikatis, V. (2015) Design of computational retrobiosynthesis tools for the design of de novo synthetic pathways, *Curr Opin Chem Biol*, vol. 28: 99-104

Manuscripts submitted

Hadadi, N., Hafner, J., Soh, K.C., Hatzimanikatis, V."Reconstruction of biological pathways and metabolic networks from *in silico* labelled metabolites", *submitted*
Seijo, M., **Hadadi, N**., Soh, K.C., Hatzimanikatis V. "BridgIt: A novel framework for bridging knowledge gaps in genome scale models." *submitted*
Hadadi, N., Ataman, M. and Hatzimanikatis, V. ''The design of "*Super E. coli*" through automated reaction network generation and genome scale models"**,** *submitted*

INTERESTS

Sport: Hiking, **Reading:** History, **Cooking:** International dishes