EPFL

Thèse n° 8382

Algorithms and flowchart for the design of synthetic biochemical networks

Présentée le 14 décembre 2020

à la Faculté des sciences de base
Laboratoire de biotechnologie computationnelle des systèmes
Programme doctoral en chimie et génie chimique

pour l'obtention du grade de Docteur ès Sciences

par

Homa MOHAMMADI PEYHANI

Acceptée sur proposition du jury

Prof. A.-C. Corminboeuf, présidente du jury
Prof. V. Hatzimanikatis, directeur de thèse
Prof. N. Kyrpides, rapporteur
Prof. J.-L. Reymond , rapporteur
Prof. B. Correia, rapporteur

École polytechnique fédérale de Lausanne

2020

"To all of the unknown people out there in any walk of life who have gambled their fortunes, their careers, and their reputations to take a risk but, in the end, failed. I'd just like to say that they should remember that it's the struggle itself that is its own reward, and the satisfaction that you knew that you gave everything you had to make the world a better place"

Nobel Lecture, 2014
by Eric Betzig

Acknowledgements

This thesis presents fairly a scientific summery of my research in the last four years. I would like to take this chance and appreciate all the individuals who directly or indirectly had positive impact on my PhD. I was very fortunate to work with many people who supported me in every single step of this long journey.

First and foremost, I would like to express my deepest gratitude to Prof. Vassily Hatzimanikatis, my thesis director. Vassily, I would like to thank you for putting your trust in me and accepting me in your laboratory. During these years, your support always gave me the motivation to overcome the inevitable difficulties that are part of any research. During stressful days of writing thesis and preparing for the PhD defense, I was calm, and it was coming from a simple sentence that you told me: "I'm not worried about your thesis". You helped me to become the scientist that I am today and for this I will be always thankful.

I also want to appreciate Prof. Irina Borodina and Prof. Christina Smolke, for the exciting research collaborations, many good ideas and interesting discussions.

I am very thankful to Prof. Clémence Corminboeuf, for accepting to preside my thesis jury, and the other members of thesis committee, Prof. Bruno Correia, Prof. Nikos Kyrpides and Prof. Jean-Louis Reymond for the time and effort they have spent on my work, and the fruitful discussion we had.

I couldn't finish my PhD without the support and guidance of one individual, Dr. Noushin Hadadi. Noushin has been more than just my postdoc and colleague, she took me under her wings during the first years of my PhD and learned me how to deal with new scientific challenges. Also, I would like to say a big thank to Dr. Ljubisa Miskovic, for all his scientific and moral support during the past four years.

I want to thank my coworkers, in particular Dr. Jasmin Hafner, Anastasia Sveshnikova, and Dr. Anush Chiappino-Pepe. Not only I learned how to do research with you, but also how to maintain motivation and momentum. It has been a pleasure to work with you, girls!

I would like to express my appreciation to the other past and present members of LCSB, for the good and collaborative atmosphere: To Milenko, Tuure, Sophia, Maria, Pierre, Robin, Daniel, Meriç., Tiziano, Yves, Vikash, Omid, Evangelia, Liliana, Joana, Asli and Aarti. Without you, all these years would have meant much less to me. I should also mention the master students that contributed to the lab atmosphere over the years:

Thomas, Cecilia, Maxime, and Remi with a special thanks to all the bright and amazing students I supervised, Kiana, Victor, Alan and Benjamin, for many hours spent in the lab.

I am grateful to Christine Kupper and Anne-Lene Odegaard for their advice, helps and administrative supports in so many occasions.

I would like to say a big thank to all my close friends in Lausanne. Zhaleh, you are the first person that I met in Lausanne, you are a wonderful friend, I will never forget all days that we spent together. My friends, Armita, Niloofar, Bahareh, Fatemeh, Mohammad, Parimah, Atena, Ashkan, Aida, Hesam, Abolfazl and Ehsan, thank you for filling these years with wonderful experiences. Your friendship means a lot to me.

I feel enormously indebted to my family, I would like to thank them by all my heart. Above all, I thank my parents, Goli and Alireza, for all the years raising me with so much love and value. I learned my first science lessons from you, I learned to work hard and honest. You have been the greatest and most influential teachers in my life. I don't have words to describe what you gave me, I want to let you know I am indefinitely grateful for everything.

I want to thank my brother and his family, Hamed, Negin, Mehrsa and Elsa, whom I miss every day. Thank you for all your encouragements, life advises, happy photos and smiles. Thousands of kilometers couldn't take us apart! I also thank my grandparents, my aunt, uncles, and cousins for always being there and their love. Also, I would like to thank my parents in law, Marziyeh and Saeed, for their acceptance, support and love.

Finally, I want to thank Amir Hossein, my biggest supporter, my best friend and my husband. Thank you for going through this journey with me, for picking me up whenever I fell down, for being the most patient person that I know, for going through all the frustrations with me and for teaching me how to ski! When you are by my side, I feel there is nothing in this world that I cannot accomplish. There are no words that can express my gratitude for all you've done and for always trying to see the best of me. I love you more and more each day and I look forward to the next steps we will make together in life.

Abstract

New drugs are needed to assure effective therapies for previously untreated diseases, emerging diseases, and personalized medicine, but the process of drug development is complex, costly, and time-consuming. This is especially problematic considering that 90% of drug candidates in clinical trials are discarded due to unexpected toxicity or other secondary effects. This inefficiency threatens our health care system and economy. Despite the advances in the cellular metabolism, our knowledge of the mechanisms governing enzymatic biotransformations in cells is far from complete, in particular regarding degradation pathways, mode of action, or side effects of drugs. Examining the mechanisms of enzymatic reactions at the cellular scale could improve our fundamental understanding of their catalytic capability, and facilitate identifying and filling the knowledge gaps. The scale and the complexity of metabolic data is ever-expanding, requiring scientists to apply more advanced computational methods to systematically store, explore, and interpret the enzymatic potential of cells.

The first step toward simulating enzymes *in silico* is to learn from their biochemical reactions in nature. To do this, we use distilled knowledge of known biochemistry in the form of generalized enzymatic reaction rules. Enzymatic rules are mathematical representations of enzymatic action mimicking the catalytic function of enzymes. They are formulated in a less specific manner (more promiscuous) to act on a broad range of substrates. In addition to reconstructing known biochemistry, the application of these reaction rules paves the way toward the discovery of novel enzymatic interactions.

In this thesis, I developed computational models, tools, and methodologies to facilitate the study of metabolism and catalytic action of enzymes. We analyzed different aspects of metabolism through five distinct studies: In a first study, in order to provide a holistic view of currently known biochemistry, we gathered biochemical data from 14 sources, covering the known metabolic networks of all species. We integrated all biological data into a high-performance database based on ontology, named LCSB DB. We further expanded the scope of LCSB DB to cover all bioactive and chemicals. LCSB DB offers fast and efficient searching of biochemical data and serves as a platform for sharing, storing, and analyzing biochemical data. In a second study, we used enzymatic reaction rules to predict all theoretically possible metabolic reactions between biological and bioactive compounds in LCSB DB. In a third study, we developed a method to find enzymes are able to catalyze orphan and predicted reactions, called BridgIT. BridgIT uses the knowledge of reactive sites on substrates to find the most similar, known biochemical reactions. We then validated the

utility of BridgIT in enzyme discovery for the design of de novo synthetic pathways producing tetrahydropalmatine and adipic acid. In the last study, we propose a workflow for rational drug design and systems-level analysis of drug metabolism, called NICEdrug.ch. NICEdrug.ch allows large-scale computational analysis of drug biochemistry (metabolic precursors or prodrugs and metabolic fate or degradation), enzymatic targets, and toxicity in the context of cellular metabolism. Finally, in the conclusion chapter, we discuss the contribution and the potential further applications of the computational tools that were developed in this thesis.

Keywords

Metabolism, biological data, systems biology, metabolic engineering, computational biology, de novo pathway design, enzyme promiscuity, enzyme annotation, drug discovery

Résumé

De nouveaux médicaments sont nécessaires pour assurer des thérapies efficaces contre des maladies non traitées auparavant, des maladies émergentes et une médecine personnalisée, mais le processus de développement de nouveaux médicaments est complexe, coûteux et prend du temps. Ceci est particulièrement problématique étant donné qu'environ 90% des médicaments testés dans des essais cliniques sont rejetés en raison d'une toxicité inattendue ou d'autres effets secondaires. Cette inefficacité menace notre système de soins de santé et notre économie. Malgré les progrès dans le domaine de du métabolisme cellulaire, notre connaissance des mécanismes gouvernants les biotransformations enzymatiques dans les cellules est loin d'être complète, comme les voies de dégradation, les modes d'action ou les effets secondaires des médicaments. L'examen des mécanismes des réactions enzymatiques à l'échelle de la cellule entière pourrait améliorer notre compréhension fondamentale de la capacité catalytique des enzymes, et faciliter l'identification et le comblement de nos lacunes dans nos connaissances. L'échelle et la complexité des données métaboliques sont en constante expansion, obligeant les scientifiques à appliquer des méthodes de calcul plus avancées pour stocker, explorer et interpréter systématiquement le potentiel enzymatique des cellules.

La première étape vers la simulation des enzymes *in silico* est d'apprendre des réactions biochimiques en nature. Dans ce travail, nous utilisons des connaissances distillées de la biochimie connue dans les règles de réaction enzymatiques dites généralisées. Les règles enzymatiques sont des modèles mathématiques imitant la fonction catalytique des enzymes et sont formulées de manière moins spécifique (et donc plus promiscuité) pour agir sur une plus large gamme de substrats. Ainsi, l'application de ces règles de réaction en plus de reconstruire la biochimie connue ouvre la voie à la découverte de nouvelles interactions enzymatiques.

Dans cette thèse, nous avons développé des modèles informatiques, des outils et des méthodologies pour faciliter l'étude du métabolisme et de l'action catalytique des enzymes. Nous avons analysé différents aspects du métabolisme à travers cinq études distinctes. Dans la première étude, afin de fournir une vue holistique de la biochimie actuellement connue, nous avons rassemblé des données biochimiques de 14 sources, couvrant les réseaux métaboliques connus de toutes les espèces. Nous avons intégré toutes les données biologiques dans une base de données haute performance basée sur l'ontologie, nommée LCSB DB. En outre, nous avons élargi la portée de LCSB DB pour couvrir tous les produits bioactifs et chimiques. LCSB DB offre

une recherche rapide et efficace des données biochimiques et sert actuellement de plate-forme standard pour le partage, le stockage et l'analyse des données dans LCSB. Dans la deuxième étude, nous nous sommes appuyés sur les données biologiques de LCSB DB. En utilisant des règles de réaction enzymatique, nous avons prédit toutes les réactions métaboliques théoriquement possibles parmis les composés biologiques. Dans la troisième étude, nous avons répondu à la question de savoir quelles enzymes sont capables de catalyser les réactions prédites. Nous avons développé BridgIT, une méthode d'annotation enzymatique qui utilise la connaissance des sites réactifs des substrats. Dans l'étude suivante, nous avons validé l'utilité de BridgIT dans la découverte d'enzymes pour la conception de voies de synthèse de novo produisant de la tétrahydropalmatine et de l'acide adipique. Dans la dernière étude, nous proposons un flux de travaux pour la conception rationnelle des médicaments et l'analyse des systèmes du métabolisme des médicaments, appelé NICEdrug.ch. NICEdrug.ch permet une analyse informatique systématique et à grande échelle de la biochimie des médicaments (précurseurs ou promédicaments métaboliques et devenir ou dégradation métabolique), des cibles enzymatiques et de la toxicité dans le contexte du métabolisme cellulaire. En conclusion, nous discutons de la contribution et des applications potentielles de cette thèse pour l'analyse à grande échelle et systématique du métabolisme.

Mots clés

Métabolisme, données biologiques, biologie des systèmes, génie métabolique, biologie computationnelle, conception de voie de novo, promiscuité enzymatique, annotation enzymatique, découverte de médicaments

Contents

List of Figures

List of Tables

List of Equations

List of Abbreviations

BNICE.ch	Biochemical Network Integrated Computational Explorer
DB	Database
GEM	Genome-scale metabolic model
LCSB	Laboratory of Computational Systems Biotechnology
SMILES	The simplified molecular-input line-entry system
EC	Enzyme Commission number
GCM	Group contribution method

Chapter 1 **Introduction**

"The first step in solving a problem is to recognize that it does exist"

Zig Ziglar

1.1 Motivation

New diseases are still being discovered. SARS-CoV-2 is responsible for the on-going severe public health and economic crisis, and the death of almost one million people (as of today, September 15 [1]) and there is currently no confirmed treatment for it.

SARS-CoV-2, a new coronavirus, attacks cells in a novel way and causes the COVID-19 disease. Our knowledge about this new virus and its mechanisms is limited to a few months of research after its appearance in December 2019 and a reason to why no drug was available to block new evolving disease. Understanding weak spots of the virus and developing a drug to target them usually needs years of research and experimentation [2]. However, in the case of coronavirus and the impending global health crisis, there is a strong social and economical pressure to shorten this time span while "social distancing", safety, sanitization and lockdowns, will only buy us the critical time to find an effective drug. As a result, the healthcare system is more than ever under the pressure to cope with the rise of new diseases and infections such as COVID-19 as well as the ongoing struggle against drug resistance pathogens [3], [4]. The costs related to healthcare is large burden for individuals and societies [5]. Despite this, a growing percentage of people experience drugs that don't benefit them. Most of the currently used drugs are approved in clinical trials which are tested on large cohorts of people based on population averages and "one-size-fits-all" mentality [3]. However, this approach neglects the fact that each person is genetically unique, lived in different environments and, as recently has been shown, has different "gut microbiome" which effects drug response [3]. Therefore, we need a major shift in the drug discovery process to reduce the time and cost of drug development, overcome drug resistance and develop personalized treatments.

Application of cutting-edge measurement and identification technologies used in biological systems resulted in large volumes of data in different levels of biology (also known as "omics data"), that are aimed to revolutionize drug design and development [3]. Examples are (i) terabits of recorded high-resolution videos

showing how cells grow, interact and divide. (ii) New sequencing technologies that have revealed the genetic code of thousands of organisms. (iii) The transcriptomic data that has provided the opportunity to study gene variability under different circumstances. (vi) Large scale proteomics data catalogued in databases enables analyzing the structure and function of proteins [6]. (v) metabolomics, the newest category of omics data, that might be the one that best represents phenotypes [6]. Even though the transcription of DNA to RNA and translation to proteins are extremely crucial, metabolites are the end product of biological systems in response to genetic and environmental perturbations[7]. For example in the field of immunology, genetic information helps to identify the risk of diseases, while metabolism determines the result of combining genetic information and environmental factors, and eventually what is the cause and manifestation of a disease [6]. Therefore, metabolism can unlock the entire disease phenotypes, and drives progress toward precision medicine [8] by offering a blueprint of cellular biochemical activities.

1.2 Metabolism

Metabolism is broadly defined as the sum of all biochemical processes that occur in organisms to maintain life. Metabolism is dauntingly large and complex; hundreds of thousands of molecules participate in thousands of metabolic reactions and altogether form the great network of metabolism that is hierarchically organized. Metabolic networks consist of smaller modules, called metabolic pathways, linking metabolites from different parts of the network to one another. Each metabolic pathway is the sequence of several metabolic reactions working together to convert a set of metabolites to products that are used by other pathways. Almost all metabolic reactions are enzymatic reactions(Figure 1.1).

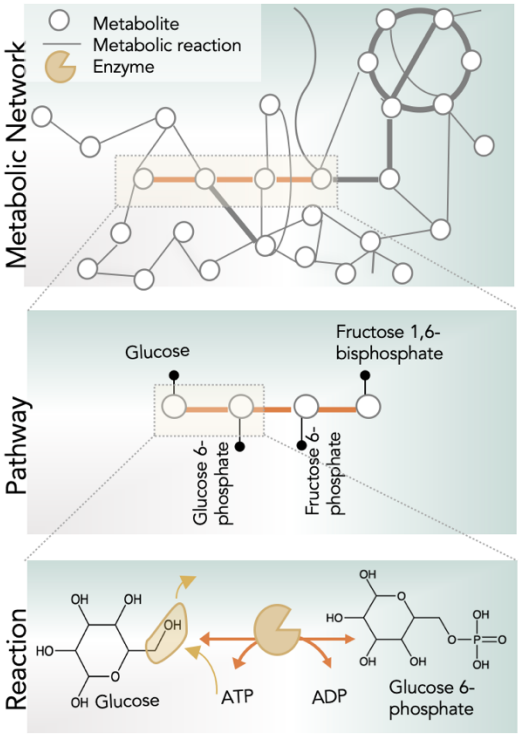


Figure 1.1: Life processes are hierarchically organized and heavily interlinked. Metabolism is the biggest biological network that is generally modeled as metabolic networks. Metabolic networks are a collection of metabolic pathways that are made up of sequential biochemical reactions. Each reaction comprised of a set of metabolites being catalyzed by an enzyme. In atomic resolution, enzymes bind to substrates and speed up the mechanism of bond-breakage-formation during metabolic reaction.

Enzymes make the transformation of metabolites occur in biologically relevant conditions in milliseconds [9]. The catalytic power of enzymes can be illustrated by evaluating how slowly metabolic reactions occur in the absence of enzymes. For example, in 2008, Lewis and Wolfenden reported that the half time – i.e., the time is needed to consume half of the substrate – for an essential reaction in the biosynthesis of hemoglobin, without an enzyme called *uroporphyrinogen decarboxylase,* it would take up to 2.3 billion years (the half of the estimated age of earth, 4.5 billion years [10]) to complete [11]! With enzymes this happens in milliseconds. Without enzymes there is no life!

Today, more than 5k enzymes are discovered and cataloged in KEGG database (Kyoto Encyclopedia of Genes and Genomes [12]), enlisting more than 10k reactions involving over 7k compounds [12]. Enzymes depending on the reaction and the substrate that they catalyze have common names with the suffix *-ase* (e.g. glucose oxidase) or *-in* (e.g. trypsin). The common names of enzymes provide little information about the nature of reactions. In addition, sometimes a certain enzyme could become known with several different names or worst, even sometimes one name is used to refer to more than one enzyme [9]. Considering the growing number of enzymes being discovered and the complexity and inconsistency of their naming, the International Union of Biochemistry assigned the Enzyme Commission (EC) number to each enzyme. EC number consists of four digits (a.b.c.d) organizing enzymes in standard classes. The first digit of EC explains the type of reaction, the second digit represents the functional group enzyme acts on, the third part stands for involved cofactors and the last digit is specific to the substrate being catalyzed [9]. EC numbers provide a basis for systematic analysis of the enzymatic reactions by computational tools used in this thesis that will be discussed in the next chapters.

1.3 Darker side of enzyme specificity

An enzymatic reaction occurs when substrate(s) binds at enzyme's active site to form the substrate-enzyme complex. The fact that the structure of the enzyme's active site is complementary to the shape of the reactive site on the substrates, ensures proper binding holds in the enzyme-substrate complex. The *lock and key* theory is one of the models often used to explain this complementary binding, wherein the enzyme binding pocket is the "lock" and the substate is the "key". Although enzymes are known to catalyse their specific substrates, many of them show promiscuous activity. The promiscuity of enzymes is referred to as the darker side of enzyme specificity and is defined as the secondary functions of enzymes which makes them able to catalyze side reactions [13], [14]. Enzymatic promiscuity can be classified into three groups based on their mechanism: (i) substrate promiscuity, (ii) catalytic promiscuity, and (iii) conditional promiscuity [15]. The substrate promiscuity refers to catalyzation of a set of diverse substrates by the same enzyme. An example of a substrate promiscuous enzyme is TP53-induced glycolysis and apoptosis regulator (TIGAR), which shows phosphatase activity on several substrates, such as 2,3-bisphosphoglycerate, 2-phosphoglycerate and fructose 2,6-bisphosphate [16]. If an enzyme catalyzes different metabolic reactions, it's called catalytic

promiscuity. The cytosine-methyltransferases that catalyzes both cytosine-methylation and cytosine-deamination is an example for this type of promiscuity. Conditional promiscuity is dependent on some environmental changes or stresses such as increasing concentration of native substrate's analogs with lower affinity to the active site [15]. It has been speculated that bacteria and archaea developed promiscuous enzymes to cope with the environmental changes, allowing them to alter and reprogram their metabolic pathways and survive in extreme conditions [17]. The application of enzyme promiscuity has motivated researchers to enhance the activity of existing enzymes and suggest new biosynthesis pathways toward more sustainable approaches in chemical industries [18].

Understanding how enzymes catalyze complex biotransformations at the atomic levels, with high specificity and efficiency is a fundamental question in biochemistry. Large scale experimental data is required to explore the catalytic potential of all enzymes and understand the underlying mechanism of their action, however, this is not practical due to time, cost, and technological limits. Therefore, computational approaches are the key to propose hypotheses, to make strategy and interpret experimental results. In this thesis, we use computational tool "*BNICE.ch*", and its enzymatic reaction rules to discover the mechanism of enzymatic reactions.

1.4 Modeling enzymes *in-silico*

Numerous computational methods have been developed to model enzymatic interactions in different levels of substrate-enzyme complex and protein-protein interactions, in different scopes ranging from single enzyme to network of enzymatic interactions of organisms for various applications. A group of methods distil the enzymatic promiscuity and formulate them in so-called generalized enzymatic reaction rules. The generalized enzymatic reaction rules simulate the activity of actual enzymes *in-silico*. Reaction rules are designed based on our knowledge about enzymatic reactions, functions and mechanisms in biochemical databases. The idea in developing enzymatic rules is to group similar metabolic reactions and formulate their substrates and biochemical activity using automatic or manual pipelines. Automatic approaches [19]–[22] are able to interpret the vast amounts of data in a short time, however, depending on the source of data and applied methods, the quality of reaction rules varies. On the other side, development of expert curated reaction rules [23], [24] (manual approach) is very time consuming but have high quality and explain precise biochemistry which is very difficult to capture correctly with automatic approaches.

BNICE.ch (Biochemical Network Integrated Computational Explorer)[23], [25], [26] developed by Hatzimanikatis et.al. in 2005, introduced the concept of enzymatic reaction rules and their applications in predictive biochemistry and *de novo* pathway design. Later, several similar methods adopted the concepts of enzymatic reaction rule [20], [27]–[30]. The database of reaction rules is the heart of BNICE.ch, where the knowledge about biochemistry is collected, manually curated, digitalized, and stored in a few hundred

electron-bond-matrix representations. Each reaction rule consists of three parts, the first part describes the atoms of reactive site, the second part formulates the atom-bond configuration inside reactive sites, and the last part explains bond breakage-formation during biotransformation [31]. The BNICE.ch reaction rules follow the same logic of classification as the first three digit of EC numbers, the last digit which is specific to a substrate is relaxed, making reaction rules more general and promiscuous. Therefore, the same rule can recognize its reactive site on a broader range of compounds, and transform them according to the mechanism of the rule to products. The structure of generated products will be searched in compound databases integrated in BNICE.ch to be identified. As the source for compounds, BNICE.ch uses the LCSB database which is a union of all reported biological, bio-active and chemical molecules in more than 14 well-known repositories. Therefore, BNICE.ch using generalized reaction rules is not only able to reconstruct known metabolic reactions but also to predict novel biotransformations. The novel reactions predicted by BNICE.ch demonstrates its predictive power to fill the knowledge gaps in metabolic networks and also its application in metabolic engineering where the discovery of *de novo* pathways is of great interest (Figure 1.2). In chapter 3 of this thesis, we demonstrate how BNICE.ch systematically explores and expands the horizon of biochemistry.

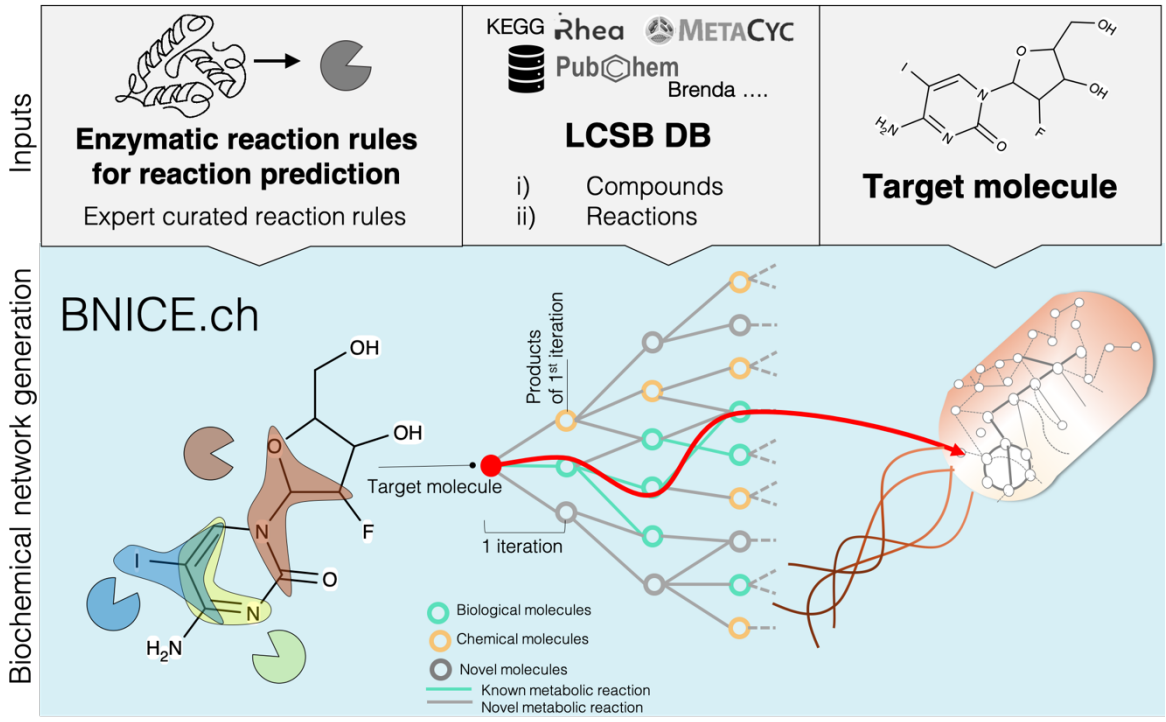


Figure 1.2: Schematic workflow of BNICE.ch illustrates the retro biosynthesis logic. This means that it backtracks the enzymic steps to find a sources compound in the organism that can generate the target molecule. Network generation starts by applying iteratively reaction rules on the target compound. In the first iteration, reaction rules scan structure of target molecule to find putative reactive sites. Then, reaction rules apply related bond-breakage-formation on the recognized reactive sites in order to generate all possible product molecules. The products of the first iteration are then used as substrates in a second iteration of reaction generation, and so on, until a complete biochemical network around the target compound is generated and a suitable precursor compound is hit. The

resulting network contains known and novel reactions. Each iteration can theoretically result in all possible chemical structures, based on the reaction rules. In order to avoid a network explosion, depending on the application, we only allow compounds to be produced that are part of the predefined search space (database membership (LCSB DB), constraints on molecular formula, molecular weight, etc.).

1.5 Enzyme annotation

New advances in computational biology has resulted in an increasing number of novel metabolic reactions predicted by tools such as BNICE.ch. However mapping back these reactions to known biochemistry and finding candidate enzymes for their catalyzation is challenging [31]. The enzyme prediction tools compare the structural similarity of novel reaction with all the known enzymatic reactions. Inspired by the theory of lock and key, they assume if the overall structure of two substrates are similar enough, most likely they can be catalyzed by the same enzyme. However in reality, only the substructure in and around reactive site is required to predict putative enzymes for novel reactions. The in-house developed computational tool, BridgIT, takes advantage of reactive site information encoded in reaction rules and instead of exploring blindly, it focuses on the reactive site and its neighborhoods in structural similarity calculations [31]. The development of BridgIT method and its applications are discussed in detail in chapters 4 and 5.

1.6 In this Thesis

In different chapters of this thesis, we use available bio-informatics and chem-informatics methods as well as novel computational tools, to learn from metabolism, predict missing pieces and develop strategies to tackle current challenges in metabolism (

Figure 1.3). In chapter 2, we review the accumulated biological data across different databases, we discuss what aspects of biochemistry they have covered, how they overlap, and how consistent they are. We also study the development of biological resources over time. In addition, we explain their differences in objects, scopes and applications, which resulted in arising of heterogeneous blocks of data. We borrow ontological database design from computer science to address this miscellany and offer a unified, curated resource for biochemical data (LCSB DB). In chapter 3, we show systematic mining of the mechanisms and function of enzymatic reactions helps to uncover their potential to catalyze other metabolites and even predict novel biochemical reactions. The novel reactions are stored in a database called ATLAS of biochemistry and offers scientists a unique resource to gather knowledge and hypotheses on the biochemistry around specific pathways. Application of novel ATLAS reactions pave the way for integrating chemicals into biochemical pathways. In chapter 4, we discuss the most important question about novel reactions: “which enzymes are able to catalyse them?” (Commonly known as enzyme annotation). The interest for enzyme annotation however is not limited to novel bio-transformations. It also covers a group of characterized reactions without

information on their associated enzymes, i.e., orphan reactions. We addressed the enzyme discovery quest, by developing a computational tool named BridgIT. BridgIT is inspired by the theory of lock and key, assuming two similar reactions will be catalysed by the same enzyme. Its novelty lies in the fact that, BridgIT inserts the information about enzyme binding pocket into reaction similarity evaluations. In chapter 5, we use BridgIT to answer practical questions in metabolic engineering. We discuss the potentials, performance and efficiency as well as the limitations of using BridgIT via two case studies in collaboration with experimental groups.

In chapter 6, we come back to our main motivation (novel methods for drug design) where we put all these pieces together to develop a workflow for drug discovery, named NICEdrug.ch. NICEdrug.ch is implemented in an open-access platform, which we aim to serve as the first resource that (1) enables a comprehensive systems-level and systematic analysis of drug metabolism, and (2) provides predictive insights to assist in rational drug design with unparalleled speed and precision, and at an unprecedented scale. As the proof-of-principle demonstration, we applied NICEdrug.ch to discover and evaluate COVID-19 drug targets and repurpose drugs. In addition, we explored the metabolic fate and toxicity of a cancer drug. As another case study, we identified new drug targets against the malaria parasite. In all these proof-of-principle studies we suggest hundreds of approved and non-toxic candidate drugs. The final chapter (Chapter 7) summarizes the results and proposes an outlook to future developments.

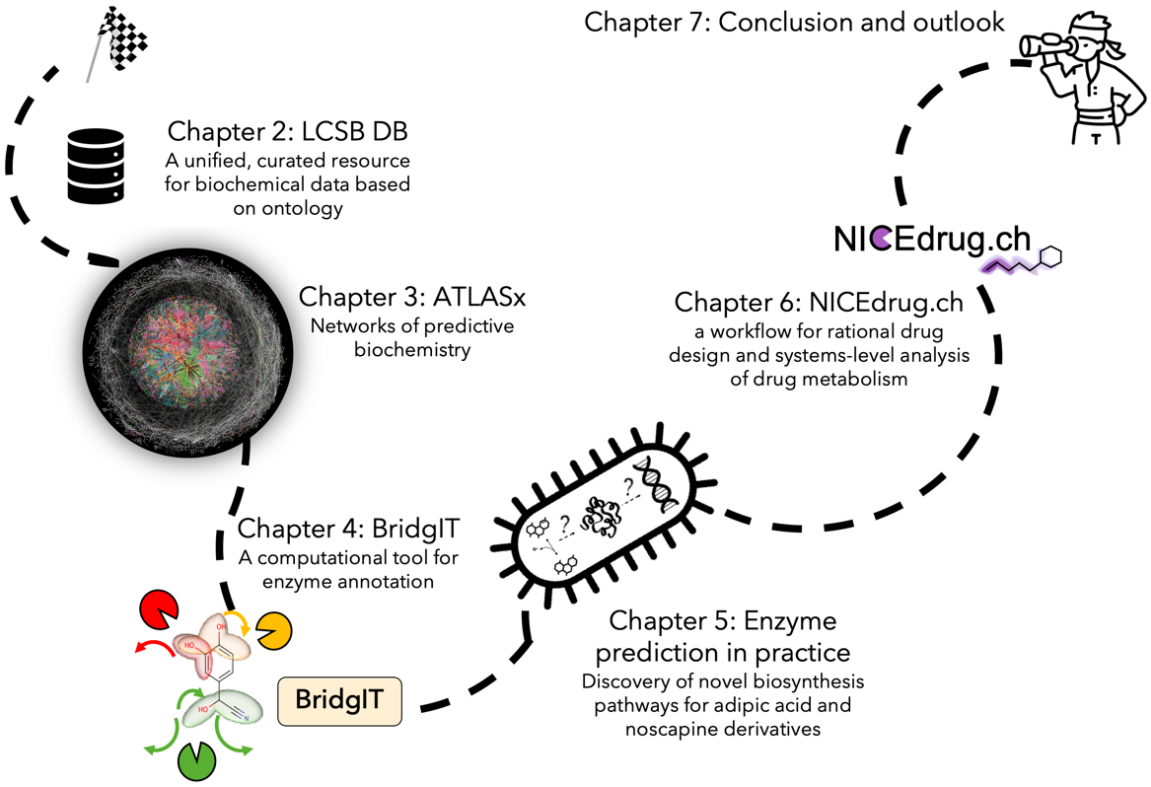


Figure 1.3: Overview on the different chapters discussed in this thesis.

1.7 Articles included in this thesis

The following list of articles is included in this thesis:

N. Hadadi[†], **H. MohammadiPeyhani**[†], L. Miskovic, M. Seijo, and V. Hatzimanikatis, "Enzyme annotation for orphan and novel reactions using knowledge of substrate reactive sites," *Proc. Natl. Acad. Sci.*, p. 201818877, Mar. 2019, doi: 10.1073/pnas.1818877116.

J. Hafner[†], **H. MohammadiPeyhani**[†], A. Sveshnikova, A. Scheidegger, and V. Hatzimanikatis, "Up-dated ATLAS of Biochemistry with new metabolites and improved enzyme prediction power," ACS Synth. Biol., May 2020, doi: 10.1021/acssynbio.0c00052.

H. MohammadiPeyhani, A. Chiappino-Pepe[†], K. Haddadi[†], J. Hafner, N. Hadadi, and V. Hatzi-manikatis, "Database for drug metabolism and comparisons, NICEdrug.ch, aids discovery and de-sign," bioRxiv, p. 2020.05.28.120782, Jan. 2020, doi: 10.1101/2020.05.28.120782 (under review).

K. Chekina[†], **H. MohammadiPeyhani**[†]**,** D. Abashkin, J. Dahlin, M. Kristensen, N. Milne, M. Sanchis, V. Hatzimanikatis, and Irina Borodina," A novel pathway for adipic acid biosynthesis in yeasts" (submitted).

J. Hafner[†], J. Payne[†], **H. MohammadiPeyhani**, V. Hatzimanikatis, and C. Smolke," A computational workflow for the expansion of heterologous biosynthetic pathways to natural product derivatives" (submitted).

H. MohammadiPeyhani[†], J. Hafner[†], A. Sveshnikova, V. Viterbo, and V. Hatzimanikatis, "ATLASx - known and predicted reactions to navigate biochemical space" (in preparation).

([†] contribuited equaly).

1.8 References

[1] E. Dong, H. Du, and L. Gardner, "An interactive web-based dashboard to track COVID-19 in real time.," *Lancet Infect. Dis.*, vol. 20, no. 5, pp. 533–534, May 2020, doi: 10.1016/S1473-3099(20)30120-1.

[2] R. McKie and S. Editor, "Coronavirus: what do scientists know about Covid-19 so far?," *The Guardian*, Apr. 30, 2020.

[3] J. Nielsen, "Systems Biology of Metabolism: A Driver for Developing Personalized and Precision Medicine," *Cell Metab.*, vol. 25, no. 3, pp. 572–579, Mar. 2017, doi: 10.1016/j.cmet.2017.02.002.

[4] E.-M. Antão and C. Wagner-Ahlfs, "[Antibiotic resistance : A challenge for society]," *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz*, vol. 61, no. 5, pp. 499–506, May 2018, doi: 10.1007/s00103-018-2726-y.

[5] E. National Academies of Sciences *et al.*, *The Affordability Conundrum*. National Academies Press (US), 2017.

[6] M. MayMay. 12, 2017, and 8:00 Am, "Big data, big picture: Metabolomics meets systems biology," *Science | AAAS*, May 12, 2017. https://www.sciencemag.org/features/2017/05/big-data-big-picture-metabolomics-meets-systems-biology (accessed Aug. 17, 2020).

[7] O. Fiehn, "Metabolomics – the link between genotypes and phenotypes," *Plant Mol. Biol.*, vol. 48, no. 1, pp. 155–171, Jan. 2002, doi: 10.1023/A:1013713905833.

[8] "Metabolomics: A Key to Realizing the Power of 'Big Data,'" *Metabolon*, Sep. 28, 2019. https://metabolon.com/metabolomics-a-key-to-realizing-the-power-of-big-data/ (accessed Aug. 17, 2020).

[9] P. K. Robinson, "Enzymes: principles and biotechnological applications," *Essays Biochem.*, vol. 59, pp. 1–41, Nov. 2015, doi: 10.1042/bse0590001.

[10] "The age of the Earth in the twentieth century: a problem (mostly) solved | Geological Society, London, Special Publications." https://sp.lyellcollection.org/content/190/1/205 (accessed Sep. 09, 2020).

[11] C. A. Lewis and R. Wolfenden, "Uroporphyrinogen decarboxylation as a benchmark for the catalytic proficiency of enzymes," *Proc. Natl. Acad. Sci.*, vol. 105, no. 45, pp. 17328–17333, Nov. 2008, doi: 10.1073/pnas.0809838105.

[12] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, "KEGG: new perspectives on genomes, pathways, diseases and drugs," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D353–D361, Jan. 2017, doi: 10.1093/nar/gkw1092.

[13] R. D. Gupta, "Recent advances in enzyme promiscuity," *Sustain. Chem. Process.*, vol. 4, no. 1, p. 2, Feb. 2016, doi: 10.1186/s40508-016-0046-9.

[14] P. Singla and R. D. Bhardwaj, "Enzyme promiscuity – A light on the 'darker' side of enzyme specificity," *Biocatal. Biotransformation*, vol. 38, no. 2, pp. 81–92, Mar. 2020, doi: 10.1080/10242422.2019.1696779.

[15] G. Piedrafita, M. A. Keller, and M. Ralser, "The Impact of Non-Enzymatic Reactions and Enzyme Promiscuity on Cellular Metabolism during (Oxidative) Stress Conditions," *Biomolecules*, vol. 5, no. 3, pp. 2101–2122, Sep. 2015, doi: 10.3390/biom5032101.

[16] I. Gerin, G. Noël, J. Bolsée, O. Haumont, E. Van Schaftingen, and G. T. Bommer, "Identification of TP53-induced glycolysis and apoptosis regulator (TIGAR) as the phosphoglycolate-independent 2,3-bisphosphoglycerate phosphatase," *Biochem. J.*, vol. 458, no. 3, pp. 439–448, Mar. 2014, doi: 10.1042/BJ20130841.

[17] M. A. Martínez-Núñez and E. Pérez-Rueda, “Do lifestyles influence the presence of promiscuous enzymes in bacteria and Archaea metabolism?,” *Sustain. Chem. Process.*, vol. 4, no. 1, p. 3, Feb. 2016, doi: 10.1186/s40508-016-0047-8.

[18] B. Arora, J. Mukherjee, and M. N. Gupta, “Enzyme promiscuity: using the dark side of enzyme specificity in white biotechnology,” *Sustain. Chem. Process.*, vol. 2, no. 1, p. 25, Dec. 2014, doi: 10.1186/s40508-014-0025-y.

[19] B. Delépine, T. Duigou, P. Carbonell, and J.-L. Faulon, “RetroPath2.0: A retrosynthesis workflow for metabolic engineers,” Jun. 2017, doi: 10.1101/141721.

[20] M. A. Campodonico, B. A. Andrews, J. A. Asenjo, B. O. Palsson, and A. M. Feist, “Generation of an atlas for commodity chemical production in Escherichia coli and a novel pathway prediction algorithm, GEM-Path,” *Metab. Eng.*, vol. 25, pp. 140–158, Sep. 2014, doi: 10.1016/j.ymben.2014.07.009.

[21] J. D. Tyzack, A. J. M. Ribeiro, N. Borkakoti, and J. M. Thornton, “Exploring Chemical Biosynthetic Design Space with Transform-MinER,” *ACS Synth. Biol.*, vol. 8, no. 11, pp. 2494–2506, Nov. 2019, doi: 10.1021/acssynbio.9b00105.

[22] T. V. Sivakumar, V. Giri, J. H. Park, T. Y. Kim, and A. Bhaduri, “ReactPRED: a tool to predict and analyze biochemical reactions,” *Bioinformatics*, vol. 32, no. 22, pp. 3522–3524, Nov. 2016, doi: 10.1093/bioinformatics/btw491.

[23] V. Hatzimanikatis, C. Li, J. A. Ionita, C. S. Henry, M. D. Jankowski, and L. J. Broadbelt, “Exploring the diversity of complex metabolic networks,” *Bioinformatics*, vol. 21, no. 8, pp. 1603–1609, Apr. 2005, doi: 10.1093/bioinformatics/bti213.

[24] J. Wicker *et al.*, “enviPath – The environmental contaminant biotransformation pathway resource,” *Nucleic Acids Res.*, vol. 44, no. Database issue, pp. D502–D508, Jan. 2016, doi: 10.1093/nar/gkv1229.

[25] K. C. Soh and V. Hatzimanikatis, “DREAMS of metabolism,” *Trends Biotechnol.*, vol. 28, no. 10, pp. 501–508, Oct. 2010, doi: 10.1016/j.tibtech.2010.07.002.

[26] N. Hadadi and V. Hatzimanikatis, “Design of computational retrobiosynthesis tools for the design of de novo synthetic pathways,” *Curr. Opin. Chem. Biol.*, vol. 28, pp. 99–104, Oct. 2015, doi: 10.1016/j.cbpa.2015.06.025.

[27] A. Cho, H. Yun, J. Park, S. Lee, and S. Park, “Prediction of novel synthetic pathways for the production of desired chemicals,” *BMC Syst. Biol.*, vol. 4, no. 1, p. 35, 2010, doi: 10.1186/1752-0509-4-35.

[28] P. Carbonell, A.-G. Planson, D. Fichera, and J.-L. Faulon, “A retrosynthetic biology approach to metabolic pathway design for therapeutic production,” *BMC Syst. Biol.*, vol. 5, no. 1, p. 122, 2011, doi: 10.1186/1752-0509-5-122.

[29] H. Yim *et al.*, “Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol,” *Nat. Chem. Biol.*, vol. 7, no. 7, pp. 445–452, May 2011, doi: 10.1038/nchembio.580.

[30] K. L. J. Prather and C. H. Martin, “De novo biosynthetic pathways: rational design of microbial chemical factories,” *Curr. Opin. Biotechnol.*, vol. 19, no. 5, pp. 468–474, Oct. 2008, doi: 10.1016/j.copbio.2008.07.009.

[31] N. Hadadi, H. MohammadiPeyhani, L. Miskovic, M. Seijo, and V. Hatzimanikatis, “Enzyme annotation for orphan and novel reactions using knowledge of substrate reactive sites,” *Proc. Natl. Acad. Sci.*, p. 201818877, Mar. 2019, doi: 10.1073/pnas.1818877116.

Chapter 2 Nature of biological data

"We are drowning in information but starved for knowledge"

John Naisbitt

2.1 Introduction

In this chapter, first we focus on data integration methods and we introduce a new approach for metabolomics data organization (LCSB DB). Then we show how proposed methods can help to gain value from quantities and diversities of data. The application of the LCSB DB will help to fill knowledge gaps in metabolic networks, to study the origins of exotic secondary metabolites with unknown biosynthesis routes and to engineer biosynthetic pathways towards chemicals of pharmaceutical or industrial interest.

The LCSB DB is the result of several years of collecting, unifying and curating biochemical data. LCSB DB integrates most of the metabolic databases with different scopes and provides a unique resource to access high quality biochemical data. Currently, LCSB DB is the backbone of many computational methods that are developed in LCSB, the same as all the methods and databases that are introduced in the next chapters of this thesis.

Since all of the presented work has been done by the author, no contribution statement was added to this chapter. The database developed here, named LCSB DB, is currently hosted on the LCSB server located in EPFL data center. The bioDB and chemDB introduced in subsection 2.3 will be used as the basis of data to create bioATLAS and chemATLAS in chapter 3.4.

2.1.1 Importance of using ontology in biology

These days by new advances in both computational and experimental technologies, a vast amount of biological data became available (Figure 2.1). Despite their usefulness, there are still several challenges working with them:

- They remain fundamentally unconnected. While sometimes there are links between entries, they are only trackable by manual browsing or through specific workflows [1].
- They are heterogeneous regarding semantics, formats, and identifiers since they follow their own standards. Integrating data among several databases is challenging.

- They typically focus on a specific topic and on a specific scale. While biological data are heavily interlinked in metabolomics, proteomics and genomics levels.
- Amounts of data are greater than to be analyzed by current infrastructures [2].

To handle this overload of heterogeneous data, high level of data organization and data integration, has become an indispensable task.

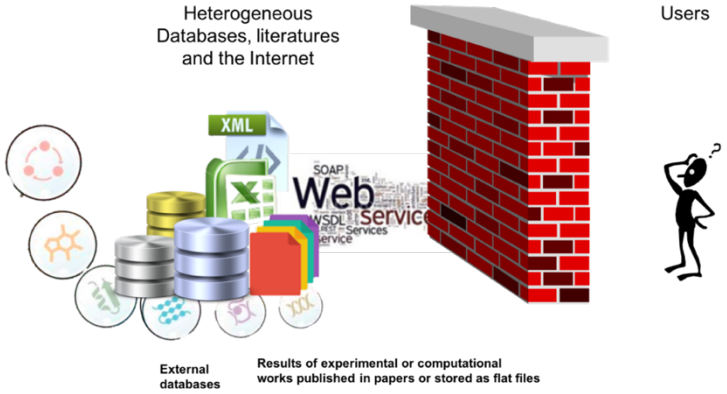


Figure 2.1: A lot of public data are available while they are heterogeneous semantics, formats, and identifiers

Ontology as a strategy for data classification was developed in computer science to facilitate data reuse and data sharing. It has been extensively used to model heterogeneous data and the reason for this success is due to its ability to keep “semantic” away from the type of data [3]. Ontological design of a database will help to organize data flexibly. Also, it will allow to abstract data and capture the relations between entries. In the chapter 2.2, we explain in details how to implement an ontology-based database.

2.2 Architecting LCSB ontological database

We aimed to design an ontological database while LCSB (Laboratory of Computational Systems Biotechnology) was using a relational database with a traditional structure (Old LCSB database). This database enclosed data in three levels of the compound, metabolic reaction, and metabolic pathway. The sources of imported data inside the database were from other external databases, or it has been generated by computational tools developed in LCSB, e.g., BNICE.ch. In the compound level, it included 16,000 compounds from KEGG 2016 database and 33 million compounds from PubChem database. In reaction level, it contained 137,000 novel reactions (generated by BNICE.ch, this set of reactions are called ATLAS reactions [4]). In pathway level, it contains the information of 1 million pathways. These data were covered inside 28 tables using MYSQL server.

In this old database structure, the number of tables will increase by bringing in new data. It is already difficult to store and analyze data in a traditional database with many tables and by considering the fast rate of data generation, the database will be drowning in data in the coming years with additional storage space and associated costs.

The idea in an ontological database is that instead of having many tables, we break down tables to "concepts". For example, in the old database we had six tables describing compounds and their properties. We abstract these tables to two concepts: "compound" and "property".

Also, we had seven tables regarding reactions. So, we defined a new concept "reaction" and interestingly, we do not need to define "property" again. But still we need to define a new concept, "relation", to cover the relation between reaction and compounds ("ownership": reaction owns compounds). In addition, to explain all the information about pathways from the old database, we just need to define "pathway" as a new concept.

These concepts together with a set of instances of each concept create a knowledge base. Instances are an extension of concepts and they preserve concept description. We decided to keep the concepts as general as possible by considering specific things as instance. The more general we are at the beginning the more flexible we will be for integrating new data and updating the existing data in future. When changing something in the level of concept, the change will be applied to all the instances of that concept automatically. For example, some of the instances stated in the concept of "property" include: "name", "chemical formula", "SMILES[1]", "energy", "error", "charge", source", "stoichiometry coefficient", "Tanimoto score" (Figure 2.2, panel A). Instances of one concept can be related to other concepts. For instance, SMILES is defined for compounds and stoichiometry coefficient is related to reaction definition. On the other hand, the name is expressible for both of them. Depending on the properties we define for each concept, instances of concepts will inherit them. The list of instances for each concept is not something fixed. So, we can add new instances when we need to cover more or when we have more information about them; it is also true about concepts. Basically, without changing the structure of the database, we can add or remove concepts or instances. That is why this approach is fully flexible with all future changes. When adding new data to the database, we just need to check if it can be expressed as an instance of defined concepts or we need to define new concepts to cover it.

When we defined the concept of "relation, we found out that as a concept, it has two levels and we could not come up with instance right after relation. For example (Figure 2.2, panel B), the relation between compound and reaction is different from reaction to compound. So, expressing ownership as an instance of relation is not enough, and we should differentiate the ownership and participation. To handle these cases,

[1] The simplified molecular-input line-entry system

we consider sub-concepts. In this case “ownership is the sub concept of relation and “owns” and “is part of” are two instances of it.

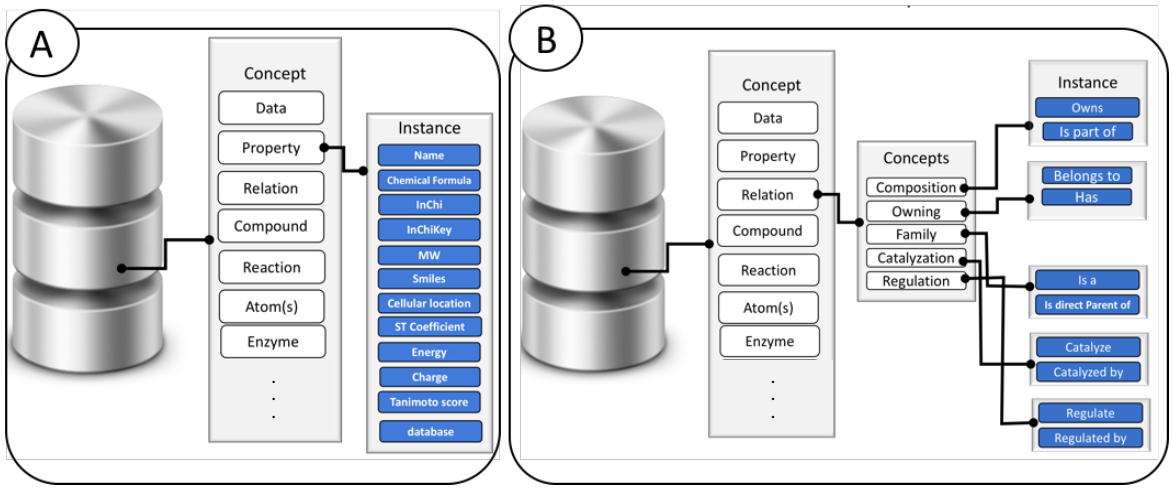


Figure 2.2: Two views of database. A) Instances defined for “property” and B) sub concepts and instances assigned to “relation”.

Finally, we store data for concepts and instances in three tables. The first table includes all the concepts (assigning concept ID to each concept). The second table, connects concepts to their instances (assigning concept ID to related instance ID), and the third table defines all the properties and relations. Basically, concepts and instance will be defined in the first two tables and the third table is linked to them. In the third table, we only need to describe how different entities are related to each other. As an example, Figure 2.3 shows how we can extract data for pyruvate from the new database. First, pyruvate is an instance of compound. So, we search for all the instances of the compound, and we filter them based on their “names” (pyruvate) which is an instance of property. In this procedure we can also extract other properties that are defined for pyruvate (Figure 2.3 panel A).To avoid replication inside database, we defined unique keys for each concept. Unique key for compounds is their canonical SMILES.

The power of ontology emerges as we look for the relations of one instance with other instances and concepts. In the example of pyruvate, we can write a simple query to find all the reactions that “own” pyruvate or pyruvate is part of them (Figure 2.3 panel B). Following the example, pyruvate participates in 191 reactions from KEGG database and 472 reactions from ATLAS of biochemistry database.

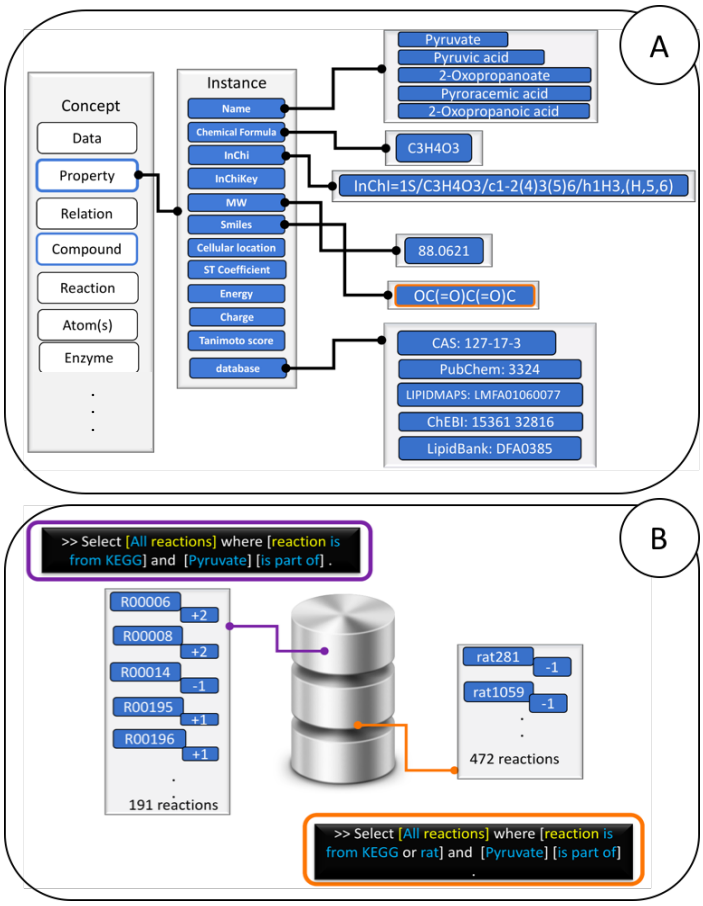


Figure 2.3 A) Extracted properties for pyruvate as an instance of compounds from the database. B) Extracted reactions which have relation with pyruvate.

To compare with what we had before as a traditional rational database, LCSB DB as an ontological database has many significant improvements.

- The number of tables from 28 (increases by bringing new data inside) decreased to 3.
- The search performance increased; due to the power of “relation” in ontology we can filter data fast and save time for collecting the same set of data from ontological database compare to a traditional, especially when dataset is large.
- LCSB DB can straightforwardly communicate with both the users and the tools.
- The new database can learn from imported data based on the relations between concepts (automated reasoning).
- While LCSB DB has three tables with a fixed structure, it has dynamic behavior, and it is extremely flexible for future updates.
- LCSB DB supports in-built methods as functions (For example: SMILES canonicalization or substructure search of compounds).

2.3 Integration of external databases

Past efforts have led to the organization of biochemical data in a range of available databases of different scopes (e.g., organism-specific vs global), curation level (e.g., manual vs automatic curation) and annotation detail. In order to reconcile the heterogenous data provided by publicly available biochemical knowledge into our database, we considered compound and reaction databases that match the biological and bioactive scope and we unified them in order to create a reference database of known biochemistry. This unified data set, named bioDB, will later be used as the starting point for the expansion of biochemical reactions.

Starting with the unification of compounds, we collected 2,297,709 compound entries from biological and bioactive databases: KEGG[5], SEED[6], HMDB[7], MetaCyc[8], MetaNetX[9], DrugBank[10], ChEBI[11], ChEMBL[12] (Table 2.1). From the collected compounds, only entries associated with a molecular structure were imported to the database. Next, the imported compounds were unified, and annotations from different databases were merged into one compound entry in the database, resulting in 1,500,222 unique 2D structural entries in bioDB. As a result of the unification procedure, a unique compound entry in bioDB can contain different resonance forms, stereoisomers, as well as dissociated and charged states of a same compound. These unification criteria are based on atoms and their connectivity in a molecule in terms of a molecular graph captured by the canonical SMILES format.

The contributions of single databases to the total of bioDB compounds varied significantly. The three biological databases KEGG compounds, SEED and MetaCyc exclusively contain biological compounds and contribute 22,447 (1.5%) compounds of bioDB. The remaining 1,477,775 (98.5%) compounds were contributed by bioactive databases which contain all compounds produced by, or known to interact with, biological systems. Remarkably, the lion's share of these bioactive compounds (1,447,079, or 97%) came from ChEMBL, suggesting that this database has the most comprehensive definition of bioactive compounds. Not only the scope of compounds was found to vary across databases, but also the ratio of unique compounds within a given database, according to our unification criteria: With 15,064 (80%) unique compounds, KEGG (biological compounds) is leading the ranking, followed by MetaCyc (12,529 unique compounds, 79%) and ChEMBL (1,365,379 unique compounds, 79%). The lowest proportion of unique compounds was detected in molecules cataloged in HMDB (43%) and drugs in KEGG (40%). This comparison illustrates the heterogeneity of database and curation standards between different resources.

We further imported all compounds from the chemical database PubChem[13] as a source of chemical identifiers (77,934,143 unique molecules). PubChem entries that could not be matched to any existing compound in bioDB were assigned to the chemical compound space (chemDB) in our database, regardless of their true origin (i.e., chemical synthesis, natural biosynthesis, or semisynthetic procedure). Some of the 77,934,143 unique compounds from the chemical space might therefore be of biological origin, but not

labeled as such, or have the potential to be derived from biological compounds in a bioengineering setting. This artificial classification of bio and chemical compounds presents the opportunity to re-assign compounds of the chemical space to the biological compound space through reaction prediction.

To create a unified reaction database, 235,698 reactions entries were collected from KEGG, BRENDA[14], Rhea[15], BiGG models[16], SEED[17], MetaNetX[9], MetaCyc[8], Reactome[18] and BKMS-react[19] (Table 2.2) and merged into 56,602 unique bioDB entries. Surprisingly, many databases contained a high number of duplicate reactions: According to our unification criteria, we observed the highest ratio of unique reactions of 94% for KEGG, followed by Model SEED with 79% unique reactions. With 22% and 31%, respectively, BRENDA and BiGG had the lowest percentage of unique reactions, indicating that many of the reactions are duplicates from a structural point of view. This quantitative assessment of reaction uniqueness further exposes the heterogenous nature of biochemical databases, and it suggests that the number of entries provided by the database hosts should be handled with care when comparing databases.

This overall unification procedure, resulted in a collection of over 1.5 million unique biological and bioactive compounds and over 56,000 unique biochemical reactions, and it provided the basis for the subsequent expansion toward hypothetical biochemistry.

The heterogeneity of the collected and unified biochemical data highlights the importance of assessing the quality and the depth of description of the metabolic reactions. We therefore checked whether or not the reactions were elementally balanced and associated to an Enzyme Commission (EC) number. We first searched for reactions containing undefined or un-processable molecular structures (e.g., polymers, proteins, compounds describing two or more disconnected structures such as salts) and other reactions that were not elementally balanced (mostly missing reaction participants, or their reaction mechanism is not known), and we found that 45% (25,296 out of 56,602) of total reactions were well-balanced (Appendix, Table 8.1). We further found that 47% (27,107) of the reactions have an EC number assigned. The highest ratio of balanced reactions with annotated EC number was found in KEGG database (80%) while BiGG models had the lowest ratio (27%).

The unification and the quality assessment of bioactive molecules and enzymatic reactions from different databases provides an overview on different resources and their curation standards, and it forms the basis for the subsequent expansion of biochemical knowledge through reaction prediction.

Table 2.1: Import and curation of compounds from different sources.

		Database	Description	Collected	Imported	Unique in source DB	
Compound DBs	**biological**	MetaCyc	Manual /Cpds of **sequenced orgs**	15,819	14,828	12,524	**Unique in LCSB DB**
		Model SEED	Manual**/KEGG** and **GSMs**	33,995	20,665	17,132	
		KEGG Comp.	Manual**/**cpds&biopolymers **relevant to biology**	18,625	17,397	15,064	
	bioactive	KEGG Drug	Manual/approved **drugs** in Japan, USA & Europe	11,140	7,766	4,514	
		Drugbank*	Approved **drugs**+discovery-phase drugs	8,350	6,279	3,850	
		ChEBI	**Chemical** Entities of **Biological Interest**	56,530	32,691	29,080	
		HMDB	Small cpds found in the **human body**	228,017	177,096	98,400	
		MetaNetX**	The metabolites in the **GSMs** + other **DBs**	200,132	183,788	87,464	
		ChEMBL	Manual /**bioactive/drug-like** cpds	1,727,112	1,595,615	1,365,379	
		Total		**2,297,709**	**2,056,125**	**1,633,407**	**1,500,222**

* Experimental drug ** not lipids cpd: compound

Table 2.2: Import and curation of reactions from different sources.

	Database	Description	Collected	Imported	Unique in source DB	
Reaction DBs	HMR	**GSMs** for **human** metabolic rxns	8,182	5,108	4,380	**Unique in LCSB DB**
	MetaCyc	Manual /rxns in pathways of **sequenced orgs**	16,052	15,438	12,726	
	KEGG	Manual/Rxns in KEGG enzyme or KEGG pathway	10,829	10,685	10,179	
	MetaNetX	The rxns in the **GSMs**+ other **DBs**	42,182	40,767	25,647	
	Reactome	Manual /reactions in **human**	1,872	1,568	814	
	Rhea	Manual curation of biochemical rxns/cpds from ChEBI	20,770	19,325	13,114	
	Model SEED	Manual/**KEGG** and **GSMs**	44,031	44,010	28,332	
	BKMS	Rxns of BRENDA, KEGG, MetaCyc & SABIO-RK	31,740	18,139	18,139	
	BiGG models	Manual/ Rxns from **GSMs**	28,299	16,581	8,681	
	Brenda	Large set of **enzyme** functional data	31,741	9,214	7,044	
	Total		**235,698**	**180,835**	**129,056**	**61,308**

GSM: Genome scale models

2.4 Interactive connection to computational tools

Today's knowledge of biochemistry does not account for the biosynthesis of many compounds that have been observed in living organisms. One of the challenging objectives in LCSB is to explore the theoretical space of biochemistry beyond experimental results by using computational methods. The Biochemical Network Integrated Computational Explorer (BNICE.ch) is an object-oriented program for predicting biotransformations and biochemically possible molecular structures [4], [20]. BNICE.ch reconstructs known reactions and predicts novel compounds and biotransformations and by applying expert-curated, generalized reaction rules on chemical compounds *in silico*. The generated results of BNICE.ch are organized in metabolic networks. The resulting networks span millions of compounds (network nodes) and biochemical reactions (network edges) connecting the compounds. The ability of "automated reasoning" in the ontology which means it can drive implicit facts from the database automatically, make it appropriate for storing and organizing such a big network and subsequently analyzing it.

On the other hand, the definition of ontology is closely related to object-oriented programming, and ontological database can constructively interact with object-oriented programs. Due to similar definition, BNICE.ch has a live connection to LCSB ontological database for getting information of compounds and reactions, discovering the link of new generated data with stored information and finally importing results to the database. Furthermore, the group contribution method used in LCSB to estimate the Gibbs free energy of reactions, is connected to LCSB DB to get the structural information of the compounds. Later, we expanded the scope of concepts in LCSB DB to integrate enzymes and their properties, which is used by BridgIT tool (chapter 4) for enzyme annotation.

Finally, LCSB DB offers for the first time a biochemical repository integrating all different levels of information about compounds, reactions, enzymes, pathways and networks, which is able to communicate flexibly with both users and computational tools.

2.5 References

[1] N. Swainston *et al.*, "biochem4j: Integrated and extensible biochemical knowledge through graph databases," *PLOS ONE*, vol. 12, no. 7, p. e0179130, Jul. 2017, doi: 10.1371/journal.pone.0179130.

[2] V. Marx, "Biology: The big challenges of big data," *Nature*, vol. 498, no. 7453, pp. 255–260, Jun. 2013, doi: 10.1038/498255a.

[3] M. Agosti, F. Esposito, and C. Thanos, Eds., *Digital libraries: 6th Italian Research Conference, IRCDL 2010, Padua, Italy, January 28-29, 2010: revised selected papers*. Berlin ; New York: Springer, 2010.

[4] N. Hadadi, J. Hafner, A. Shajkofci, A. Zisaki, and V. Hatzimanikatis, “ATLAS of Biochemistry: A Repository of All Possible Biochemical Reactions for Synthetic Biology and Metabolic Engineering Studies,” *ACS Synth. Biol.*, vol. 5, no. 10, pp. 1155–1166, Oct. 2016, doi: 10.1021/acssynbio.6b00054.

[5] M. Kanehisa and S. Goto, “KEGG: kyoto encyclopedia of genes and genomes,” *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, Jan. 2000.

[6] R. Overbeek *et al.*, “The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes,” *Nucleic Acids Res.*, vol. 33, no. 17, pp. 5691–5702, Sep. 2005, doi: 10.1093/nar/gki866.

[7] D. S. Wishart *et al.*, “HMDB: the Human Metabolome Database.,” *Nucleic Acids Res.*, vol. 35, no. Database issue, pp. D521-6, Jan. 2007, doi: 10.1093/nar/gkl923.

[8] R. Caspi *et al.*, “The MetaCyc database of metabolic pathways and enzymes,” *Nucleic Acids Res.*, vol. 46, no. D1, pp. D633–D639, Jan. 2018, doi: 10.1093/nar/gkx935.

[9] S. Moretti, O. Martin, T. Van Du Tran, A. Bridge, A. Morgat, and M. Pagni, “MetaNetX/MNXref – reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks,” *Nucleic Acids Res.*, vol. 44, no. D1, pp. D523–D526, Jan. 2016, doi: 10.1093/nar/gkv1117.

[10] D. S. Wishart *et al.*, “DrugBank 5.0: A major update to the DrugBank database for 2018,” *Nucleic Acids Res.*, vol. 46, no. D1, pp. D1074–D1082, Jan. 2018, doi: 10.1093/nar/gkx1037.

[11] J. Hastings *et al.*, “ChEBI in 2016: Improved services and an expanding collection of metabolites,” *Nucleic Acids Res.*, vol. 44, no. Database issue, p. D1214, Jan. 2016, doi: 10.1093/NAR/GKV1031.

[12] A. Gaulton *et al.*, “The ChEMBL database in 2017,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D945–D954, Jan. 2017, doi: 10.1093/nar/gkw1074.

[13] S. Kim *et al.*, “PubChem 2019 update: improved access to chemical data,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D1102–D1109, Jan. 2019, doi: 10.1093/nar/gky1033.

[14] I. Schomburg, A. Chang, O. Hofmann, C. Ebeling, F. Ehrentreich, and D. Schomburg, “BRENDA: a resource for enzyme data and metabolic information,” *Trends Biochem. Sci.*, vol. 27, no. 1, pp. 54–56, Jan. 2002, doi: 10.1016/S0968-0004(01)02027-8.

[15] A. Morgat *et al.*, “Updates in Rhea--a manually curated resource of biochemical reactions.,” *Nucleic Acids Res.*, vol. 43, no. Database issue, pp. D459-64, Jan. 2015, doi: 10.1093/nar/gku961.

[16] J. Schellenberger, J. O. Park, T. M. Conrad, and B. Ø. Palsson, “BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions,” *BMC Bioinformatics*, vol. 11, no. 1, p. 213, Apr. 2010, doi: 10.1186/1471-2105-11-213.

[17] R. K. Aziz *et al.*, “SEED Servers: High-Performance Access to the SEED Genomes, Annotations, and Metabolic Models,” *PLoS ONE*, vol. 7, no. 10, p. e48053, Oct. 2012, doi: 10.1371/journal.pone.0048053.

[18] D. Croft *et al.*, “Reactome: a database of reactions, pathways and biological processes,” *Nucleic Acids Res.*, vol. 39, no. Database, pp. D691–D697, Jan. 2011, doi: 10.1093/nar/gkq1018.

[19] L. Jeske, S. Placzek, I. Schomburg, A. Chang, and D. Schomburg, “BRENDA in 2019: a European ELIXIR core data resource,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D542–D549, Jan. 2019, doi: 10.1093/nar/gky1048.

[20] K. C. Soh and V. Hatzimanikatis, “DREAMS of metabolism,” *Trends Biotechnol.*, vol. 28, no. 10, pp. 501–508, Oct. 2010, doi: 10.1016/j.tibtech.2010.07.002.

Chapter 3 **ATLASx - Databases for predictive biochemistry**

"Look deep into nature and you will understand everything better"

Albert Einstein

The first version of ATLAS of Biochemistry, published in ACS Synthetic biology in 2016 [1], was developed by Dr. Noushin Hadadi (as the leading scientist), Dr. Jasmin Hafner and Adrian Shajkofci. Later, the interest of research community encouraged us to update ATLAS database. In 2019, the updated ATLAS of Biochemistry with more compounds and improved enzyme annotation published in ACS Synthetic Biology journal [2] as a technical note. This work has been performed in collaboration with Dr. Jasmin Hafner (reconstruction of KEGG reactions, manuscript), Anastasia Sveshnikova (enzyme prediction) and Alan Scheidegger (curation of reactions). The author of this thesis has been in charge of the completion of the manuscript, reaction prediction and enzyme annotation. Dr. Jasmin Hafner and the author of this thesis equally led the project (subchapter 3.3). Furthermore, the results presented in the extended version of ATLAS (bioATLAS and chemATLAS, subchapter 3.4) have been obtained in collaboration with Dr. Jasmin Hafner (pipeline and website development, manuscript), Anastasia Sveshnikova (reactive site analysis) and Victor Viterbo (curation of reaction databases). The author of this thesis has been in charge of the manuscript, as well as data curation, reaction generation, pipeline and database development. Dr. Jasmin Hafner and the author of this thesis equally led this project. Publication related to ATLASx is under preparation (subchapter 3.4). All the mentioned projects are supervised by Prof. Vassily Hatzimanikatis.

3.1 Introduction

The availability of different omics data from genomics, transcriptomics, and metabolomics provides better observation over cellular mechanisms. To analyse and interpret the big biological data, advanced computational tools developed. However, the observed metabolic reactions involve a small portion of measured metabolites, and the biochemical function and metabolism of many metabolites are remained unknown. Therefore, our understanding of metabolism is still far from complete. The "ATLAS of Biochemistry" is an ongoing effort to explore and expand our knowledge about metabolism.

In this chapter, we first introduce the concept of biochemical knowledge gaps (Subchapter 3.1.1), then we review the current state of computational methods that are used to fill gaps in biochemical networks (Subchapter 3.1.2). Next, we discuss the ATLAS methodology (Subchapter 3.2), and we present the results of

the ATLAS update 2018 (Subchapter 3.3). Finally, we introduce extended versions of ATLAS, named ATLASx (Subchapter 3.4).

3.1.1 Knowledge gaps or dark matter in metabolism

Biological “dark matter” is an umbrella term for underground processes in biology that are difficult to measure and barely understood[3], such as weak molecular interactions, viscosity and crowding effects, post-translational modifications, non-coding RNAs[4], non-culturable microbes[5]. In metabolism, “dark matter” designates biochemical processes where knowledge is still sparse, such as underground metabolism resulting from promiscuous enzymatic activity[6], [7], undetected plant natural products and their uncharacterized biosynthesis pathways, and chemical damage of metabolites[8]. These unknowns limit our general understanding of metabolism, which is key to discover mechanisms involved in cancer[9] or vector-borne diseases[10], or to find new medicines through drug discovery from plant natural products[11]. Furthermore, these knowledge gaps hamper the advancement of bioengineering applications like the creation of sustainable cell factories for the green production of commodity chemicals and pharmaceuticals.

The ultimate approach to identify new enzymatic functions and to detect novel natural products are biochemical assays. However, given the vastness of unknown elements left to discover, it is essential to generate hypotheses on potential biochemical functions, and ultimately to guide experimental efforts. Remedy was hoped to come from genomic, transcriptomic, proteomic and metabolomic data, but linking these data to metabolic functions remains difficult[12], [13]. As an example, 25 % of proteins in *E. coli,* one of the best studied model organism, do not have a function assigned[14], and almost 10,000 metabolites are orphan in the Kyoto Encyclopedia of Genes and Genomes (KEGG)[1], meaning that they are not integrated in any biochemical reaction in KEGG. Hence, computational approaches are needed to systematically explore the metabolic dark matter arising from the elasticity of enzymatic catalysis in an unbiased and global approach.

3.1.2 Toward characterizing knowledge gaps

The past decades have shown increasing interest in computational approaches to biological questions. Diverse tools have emerged that can bridge the knowledge gaps in metabolism through cheminformatic predictions of potential metabolic reactions, uncharacterized metabolites and novel enzyme functions. Most of these tools have been developed for metabolic engineering applications, where the objective is to find biosynthesis routes that produce a given target compound in a host organism[15]–[19]. This problem is solved by biochemically “walking back”, reaction step by reaction step, from the desired target to known precursor compounds that are produced by the host organism. This procedure is called *retrobiosynthesis* and implemented in a range of tools such as BNICE.ch[20], [21], GEM-Path[22], NovoPathFinder[23], NovoStoic[24], ReactPRED[25], RetroPath[26], [27], Transform-MinER[28]. These methods rely on the

concept of *generalized enzymatic reaction rules*. A reaction rule encodes the biochemistry of a substrate-promiscuous enzyme by describing the pattern of the reactive site recognized by the enzyme, as well as the bond rearrangement performed by the enzyme on the substrate. By applying the rule on a substrate that is non-native to the represented enzyme, the rule can predict if (i) the substrate can be recognized by the enzyme, (ii) if the biotransformation can occur, and (iii) what will be the product molecule(s). The concept of reaction rules is also employed by enviPath[29], a platform for predicting biodegradation mechanisms, and by MINEs[30], a database predicting potential biological products for mass-spectrometry applications.

3.2 ATLAS of Biochemistry methodology

However, all of the named tools for predictive biochemistry are specific to a given research or engineering question. During the endeavor to address the knowledge gap in metabolism, we employed the computational tool BNICE.ch[16], [31]–[35] (Biochemical Network Integrated Computational Explorer), to explore the "reaction space" within the compounds that are known to be present in biological systems.

BNICE.ch tool consists of (i) a large set of expert-curated, generalized reaction rules and (ii) a network-generating algorithm which apples reaction rules on each compound. Basically, reaction rules distill the knowledge of substrate reactive sites in biochemistry and digitalize them in a few hundred Bond-Electron Matrix (BEM) representations. Inside each reaction rule, a BEM is used to describe the reactive site of a molecule that will be recognized by an enzyme, and a second matrix (difference BEM) describes the bonds that need to be rearranged in a molecule in order to form the product. Based on the concept of generalized enzyme reaction rules, BNICE.ch asks how we can first reconstruct known biochemistry and on top of that discover and characterize new biochemical reactions.

Traditionally, to generate ATLAS reactions using BNICE.ch, each reaction rule was applied to each compound and all the potential products were analyzed. Reactions only producing compounds that belong to the biological or biochemical compound space were imported to the database as ATLAS reactions. However, by expanding the scope of ATLAS to larger compound databases, for example PubChem which incorporate more than 70 M compounds, application of BNICE.ch on all the compounds was practically impossible. In order to handle the complexity arises due to the millions of compounds in bigger databases, we designed our pipeline into two phases. In the first phase, we scanned the compounds with the generalized reaction rules to identify those with at least one reactive site. Such pre-selected compounds have the potential to be the substrate of biochemical reactions. Moreover, with such analysis, we could identify *all* the targets of the known enzyme and therefore capture the potential space of enzyme promiscuity. In the second phase, employing BNICE.ch we applied the identified generalized reaction rules on the pre-selected compounds and we observed how many of the identified compounds can participate in a reaction. Then, if the product of the application of the generalized reaction rules on the substrate exist in the space of known compounds (chemicals and biological),

we imported reaction to ATLAS database. Each ATLAS reaction is annotated with an estimated value for the Gibbs free energy of reaction and an EC-number up to the third level. Also, to guide the further experimental implementation of the novel proposed reactions, we assigned to them the best candidate enzyme(s) using BridgIT tool.

The origin of compounds analysed by ATLAS pipeline defines the scope of ATLAS project, abbreviated with ATLASx. For example, the original ATLAS integrates compounds catalogued in KEGG database. In the next versions of ATLASx we aim to be independent of a specific compound database and we expand the scope of ATLAS to all bio and bioactive molecules (bioATLAS), and even we take a step further to extrapolate the known metabolism towards the space of chemical compounds (chemATLAS).

3.3 Update - ATLAS of biochemistry

The following subchapter presents the updated version of ATLAS of biochemistry, which published in ACS synthetic biology as a technical note. The work has been achieved in collaboration with Dr. Jasmin Hafner (reconstruction of KEGG reactions, manuscript), Anastasia Sveshnikova (enzyme prediction) and Alan Scheidegger (compilation of reactions). The author of this thesis, has been in charge of the manuscript, reaction prediction and enzyme annotation. Prof. Vassily Hatzimanikatis supervised the project as well as the completion of the manuscript.

Full list of authors: J. Hafner†, H. MohammadiPeyhani†, A. Sveshnikova, A. Scheidegger, and V. Hatzimanikatis, "Up-dated ATLAS of Biochemistry with new metabolites and improved enzyme prediction power," ACS Synth. Biol., May 2020, doi: 10.1021/acssynbio.0c00052 († contributed equally, * corresponding author).*

3.3.1 ATLAS of biochemistry over years

The original and also updated ATLAS of Biochemistry[1] present the effort to map dark matter in biochemistry by predicting novel reactions between compounds known to the KEGG compound database[36].). The utility of original ATLAS has been recognized by several reviews as a source of novel metabolic reactions for enzyme and metabolic engineering[18], [37], [38]. More recently, Yang *et al.* experimentally validated hypothetical ATLAS reactions and used them to construct novel one-carbon assimilation pathways[39]. However, ATLAS was created based on the biochemical knowledge available in KEGG 2015[36]. Since then, KEGG has added 802 new metabolites, 918 new reactions, and 633 enzymes to its collection. Here, we present an updated version of ATLAS created from KEGG 2018 using an increased set of generalized reaction rules following the same procedure as explained in section 3.2.

Updated ATLAS contains ~150,000 reactions, out of which ninety-six percent are novel. Interestingly, we found that the newly available data validated 107 novel reactions predicted in ATLAS 2015. Furthermore, we improved the accuracy of the enzymes that are predicted for catalyzing novel reactions.

In the next sections, we present detailed statistics on the updated ATLAS and highlight the improvements with regard to the original version. The updated ATLAS is available at https://lcsb-databases.epfl.ch/atlas.

3.3.2 Updated tools and methods

Since 2015, two main aspects of ATLAS workflow have been updated, which were applied to generate the updated version of ATLAS. First, the set of bidirectional reaction rules was increased from 360 to 400. Second, we applied the most recent version of BridgIT to predict putative enzymes for novel compounds, and we report the top three enzyme matches for each. The 40 new rules were created to reconstruct the exact reaction mechanism of an additional number of 510 KEGG reactions that were not considered previously (i.e., KEGG reaction R03223).

3.3.3 Overall statistics

ATLAS 2018, based on KEGG 2018, now has 149,052 reactions, out of which 5,779 are known to KEGG. Compared to 2015, we added 510 known and 11,173 novel reactions. Thanks to the predicted reactions, ATLAS now integrates 4,587 out of 9,857 disconnected, or “orphan”, KEGG metabolites, which were not participating in any known biochemical reaction.

3.3.4 Increased coverage of KEGG reactions

The KEGG database contained 18,254 compounds as of February 2018 (Table 3.1). In a first preprocessing step, we removed 999 compounds without clearly defined molecular structures (e.g., polymers, proteins). The filtered dataset comprised 17,255 compounds, out of which 9,857 were not involved in any KEGG reaction. These orphan compounds did not participate in any known biotransformation in the KEGG metabolic space.

Out of the 10,829 reactions in KEGG, 76 involved compounds with an undefined structure that were removed, resulting in a filtered set of 10,753 reactions. Out of these, 8,118 reactions were reconstructed with BNICE.ch reaction rules. We observed three different types of reaction reconstruction: 5,779 reactions were exactly reconstructed, meaning that the reactions generated by BNICE.ch use the same cofactors as in KEGG. Another 1,705 reactions were reconstructed using alternative cofactors, out which 123 reactions were poorly characterized in KEGG (i.e., reaction mechanism not known, incomplete reaction). The remaining 634 reactions were reconstructed in two (408 reactions), three (145 reactions) or four (81 reactions) consecutive reaction steps.

Table 3.1: Overview of compound, reaction, and enzyme statistics in KEGG and ATLAS.

		ATLAS 2015	ATLAS 2018	Percent change
KEGG compounds	**Total number of compounds**	**17,450**	**18,254**	**+5%**
	Filtered compounds *(fc)*	16,798	17,255	
	Orphan KEGG compounds *(okc)*	9,371 *(56% of fc)*	9,857 *(57% of fc)*	
KEGG reactions	**Total number of reactions**	**9,135**	**10,829**	**+19%**
	Filtered reactions	8,592	10,753	
BNICE.ch	**Number of bidirectional enzymatic reaction rules**	**360**	**400**	**+11%**
KEGG reaction reconstruction	**Covered reactions total**	**6,651**	**8,118**	**+22%**
	Exact coverage	5,270	5,779	
	Alternative cofactor usage	916	1,705	
	2-step reconstruction	387	408	
	3-step reconstruction	78	145	
	4-step reconstruction	-	81	
ATLAS statistics	**Total number of reactions**	**137,877**	**149,052**	**+8%**
	Novel reactions	132,607	143,272	
	Total number of compounds	**10,362**	**10,939**	
	Number of orphan compounds integrated in ATLAS	3,945 *(42% of okc)*	4,587 *(47% of okc)*	
Consistency of EC numbers *	1st level EC match	79,058	138,168	+75%
	2nd level EC match	65,854	126,689	+92%
	3rd level EC match	47,918	94,168	+96%

* Number of matches between the EC assignment from the reaction rules and the EC numbers assigned by BridgIT for novel reactions in ATLAS

A total of 2,635 KEGG reactions were not reconstructed with BNICE.ch. First, 1,546 reactions did not fulfill the BNICE.ch requirements for reconstruction, such as reactions involving polymer structures, generic compounds, or compounds without a defined molecular structure, as well as elementally unbalanced reactions and stereoisomerase reactions. Additionally, the reaction rules are organized according to the Enzyme Classification (EC) system, so each reconstructed or predicted reaction is automatically assigned a third-level EC number corresponding to the non-substrate specific EC classification of the reconstructing reaction rule. Another 308 reactions had partial or missing EC number annotations, indicating that the reaction mechanisms are not known and therefore no rule has been created for these reactions. The remaining 862 reactions were not reconstructed because their reaction mechanisms are very specific and hence not readily generalizable.

3.3.5 Predicted ATLAS reactions validated in KEGG and other databases

To validate the predicted reactions in ATLAS, we analyzed the novel reactions predicted in 2015 that became known in KEGG 2018. Out of the 958 reactions newly added to KEGG, only 239 reactions involved compounds that were already present in KEGG 2015, meaning that they could have been predicted in the original ATLAS. Out of these 239 reactions, 107 were already present in ATLAS. In other words, the existence of hypothetical reactions in ATLAS 2015 was confirmed in KEGG 2018, demonstrating the predictive power of BNICE.ch.

Next, we examined the enzymes that BridgIT suggested in ATLAS 2015 for these 107 novel reactions, out of which 75 had an enzyme assigned. Interestingly, we found that the predicted EC numbers for 64 out of 75 reactions match the EC number proposed in KEGG up to the third level. For example, the novel reaction rat104204 was predicted to have an EC number of 2.4.1.-. BridgIT suggested R08946 as the most similar reaction, which was known to be catalyzed by 2.4.1.245. In 2018, KEGG confirmed the promiscuous activity of 2.4.1.245 for this reaction and named it R11306.

In ATLAS 2018, we additionally mapped the novel reactions to reaction databases other than KEGG. Interestingly, we found that 1118 predicted reactions in ATLAS were not actually novel, but known to at least one of the repositories Brenda, Reactome, HMR, MetaCyc, MetaNetX, BIGG or Rhea, which shows that the predictive power of ATLAS goes beyond KEGG. ATLAS reactions that can be found in any of these databases are linked accordingly in the updated version.

3.3.6 Improvements in the prediction of enzymes for ATLAS reactions

To find putative enzymes for the reactions in ATLAS, we applied the enzyme prediction tool BridgIT. With the latest version of the tool, the new predictions were significantly better in the updated ATLAS: BridgIT correctly matched 92% of ATLAS reactions to the same EC class as BNICE.ch rules, whereas the previous version only matched around 60% (Table 3.1). For each ATLAS reaction, we provide the top three candidate enzymes, and we also include BridgIT results for known KEGG reactions to provide alternative enzymes for a known reaction.

As a qualitative example of an improved prediction, we analyzed the ATLAS reaction rat109456, whose closest BridgIT candidate had a low matching score of 0.67. In ATLAS 2018, the reaction is now known and BridgIT found three very similar reactions, the first of which having a higher score than in the previous version (Figure 3.1).

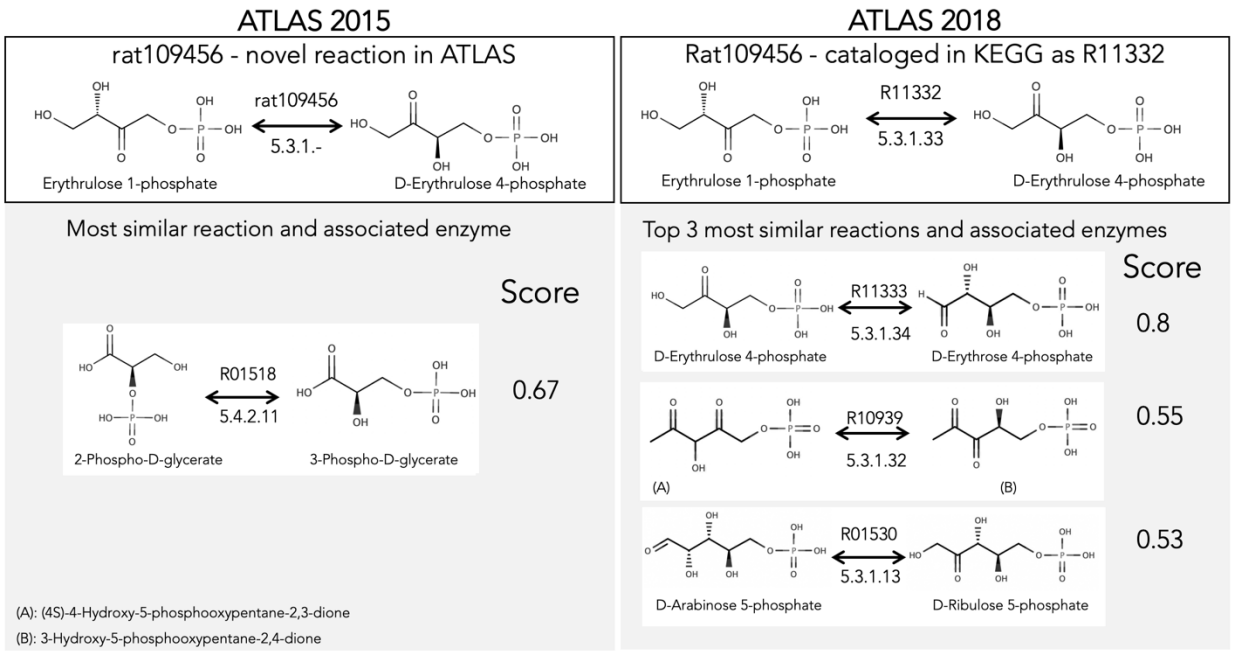


Figure 3.1: The reaction with ATLAS identifier rat109456 is an example of a reaction that was novel in ATLAS 2015 and that is now cataloged in KEGG. (left) In ATLAS 2015, the earlier version of BridgIT provided the most similar known reaction, and associated enzyme, for the ATLAS reaction with the ID. (right) In ATLAS 2018, the same reaction is now cataloged in KEGG as R11332 with EC 5.3.1.33. Other than the native enzyme with EC 5.3.1.33, BridgIT provides three alternative enzyme candidates that might also catalyze the reaction.

3.3.7 Conclusion

This study demonstrates the dynamic nature of biochemical knowledge and highlights the need for continuous updates of database-dependent applications. The updated ATLAS database contributes to fill the gaps in our current knowledge of metabolism by expanding the boundaries to novel predicted metabolic reactions. The updated ATLAS database is freely available online for academia upon request.

3.4 bioATLAS and chemATLAS - reactions emerging from biological and bioactive compounds

The following subchapter is the result of a collaborative project together with Dr. Jasmin Hafner, Anastasia Sveshnikova and Victor Viterbo. Dr. Jasmin Hafner investigated the properties of bioATLAS and chemATLAS networks (section 3.4.4) and also she implemented the online LCSB web platform for pathway search in ATLAS network, further she showed the application of pathway search in section 3.4.6. Anastasia Sveshnikova analysed the results of reaction rule assignment to the compounds in bioATLAS (section 3.4.2), also she compared pathway prediction results of ATLAS with pathways cataloged in MetaCYC database (section 3.4.5). Victor Viterbo, helped in the organization and curation of reaction databases under supervision of the author. The author of this thesis was in charge of database management and application of ATLAS pipeline in both steps of reactive site identification and reaction generation (section 3.4.1, 3.4.2 and 3.4.3). Further, she provided enzyme prediction results for gap filling application in section 3.4.6. The manuscript corresponding to this project is under preparation. Prof. Vassily Hatzimanikatis supervised the project as well as the completion of the manuscript.

Full list of authors: H. MohammadiPeyhani†, J. Hafner†, A. Sveshnikova, V. Viterbo, and V. Hatzimanikatis, "ATLASx - known and predicted reactions to navigate biochemical space" (in preparation, † contributed equally, * corresponding author)*

3.4.1 ATLASx -Networks for predictive biochemistry

One major drawback of ATLAS is its limitation to KEGG compounds. Many drugs and plant natural products with undefined or putative biological function are not part of KEGG, and therefore not included in ATLAS. Predicting enzymatic reactions from biochemical compounds retrieved from databases other than KEGG will help to integrate information from different sources, and to expand the scope of our predictions, and finally enhance the application range and the predictive power of the database.

In the following, we present ATLASx, an online biochemical resource providing reliable predictions of biochemical reactions and pathways for synthetic biologists and metabolic engineers. ATLASx unifies biochemical reactions and compounds from 14 different sources into one curated dataset, called bioDB. bioDB holds 1.5 million unique biological or bioactive compounds and 56 thousand unique biochemical reactions, forming the basis for the second achievement, the prediction of a hypothetical biochemical space (Please see subchapter 2.3 for details of data collection and curation in bioDB). Following ATLAS procedure explained in section 3.2, and by applying 490 bidirectional, generalized reaction rules from BNICE.ch on the collected biological and bioactive compound, we predicted 1.6 million potential biotransformations between bioDB compounds. Another 3.6 million reactions were found to connect bioDB compounds with molecules only found in chemical databases, adding up to a total of 5.2 million predicted reactions. From this new wealth of generated information, we characterized the connectivity and reactivity of biologically important molecules, and we showed that ATLASx pathway predictions could recover 97.5% of known biological pathways from MetaCyc. Finally, we provide access to ATLASx through an online web interface that features tools for pathway design and network exploration, which can be readily used for the design of novel metabolic pathways. The database can be accessed at https://lcsb-databases.epfl.ch/Atlas2.

3.4.2 Unification and expansion of biochemical knowledge

Biochemical knowledge is dispersed in biochemical databases of different biological scopes, varying level of information detail and diverse target applications. This situation makes it difficult to reliably detect knowledge gaps, since the missing piece of information in one database could be present in another resource. This circumstance requires the prior unification of known biochemical resources to provide a basis for the ultimate objective, the extension of known to hypothetical metabolism.

To achieve this objective, we established the following workflow (Figure 3.2): The first step (*Unification*) consisted of collecting metabolic reactions and biochemical compounds from different publicly available databases, and to merge them into a consistent and duplicate-free database, called bioDB (for more information please see subchapter 2.3). In step two (*Curation*), compounds were annotated with molecular identifiers and reactions were annotated with reaction mechanisms. In step three (*Expansion*), we applied the generalized reaction rules from BNICE.ch to the collected compounds in bioDB to generate all possible reactions producing known biological or chemical products. In the process, we reconstructed known reactions and we discovered novel, hypothetical reactions, which we stored in the ATLASx database. In step four (*Analysis*), we analyzed the connectivity of the biochemical reaction networks before and after reaction prediction, and we assessed the integration of compounds not previously connected in known biochemical networks. The results were finally made available on our website (https://lcsb-databases.epfl.ch/Atlas2) (*Distribution*).

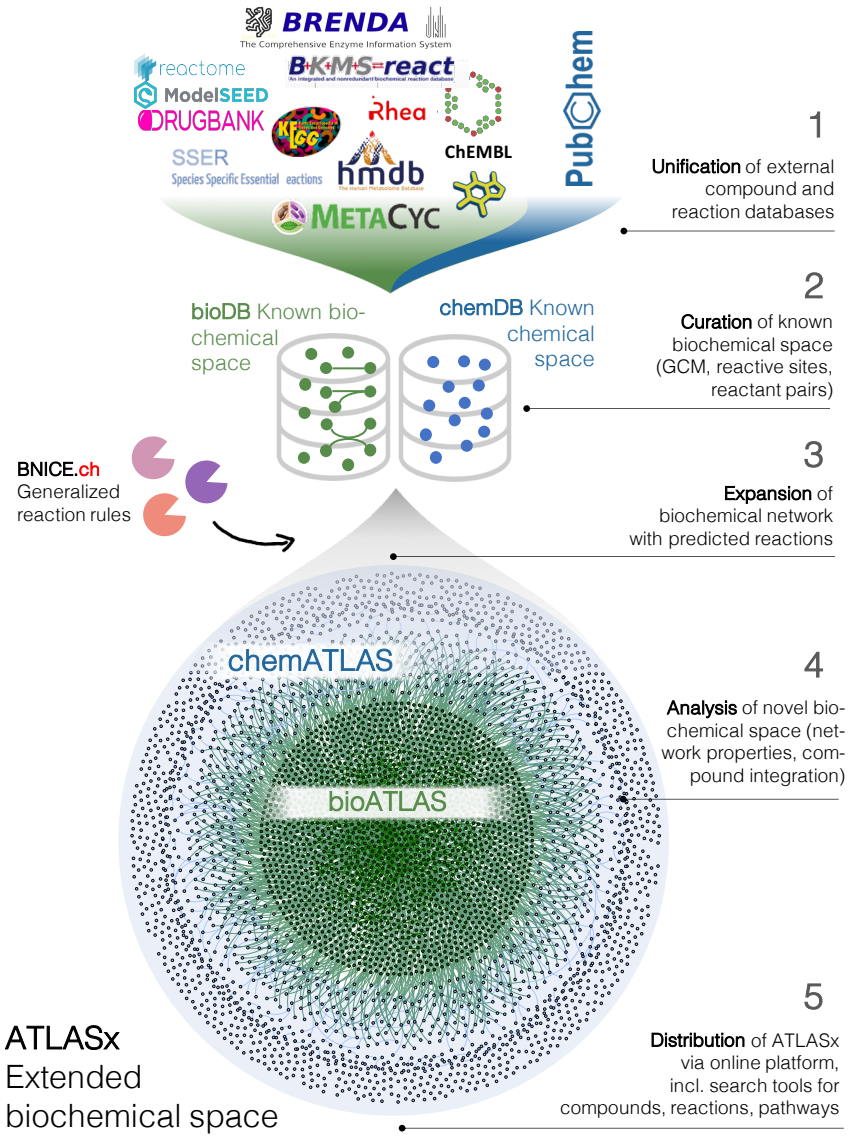


Figure 3.2: ATLAS workflow applied on the space of known biological and bioactive compounds

3.4.3 Reactive sites detected in all biological and almost all bioactive compounds

Functional groups, or reactive sites, are important features of biochemically active compounds as they designate the parts of a molecule that are recognized by enzymes and further transformed. To determine the biochemical reactivity of the collected biological and bioactive compounds, we applied the reactive site recognition encoded in the 490 BNICE.ch reaction rules to the compounds in bioDB, excluding those with more than one disjoint molecular structure (e.g. salts). The number of molecules screened for reactive sites summed up to 1,500,222 biological and bioactive compounds.

As a result of the screening for reactive sites, each compound was assigned a list of reaction rules that can recognize one or more reactive sites on the molecular structure. We found that 1,498,307 out of 1,500,222, or 99.8%, of collected biological and biochemical compounds had at least one reactive site. We found that

most of the compounds (87%) had between 50 and 200 reaction rules assigned, and contained between five to twenty carbon atoms (Figure 3.3 panel A). From the remaining 1,915 compounds without any reactive site, 958 had unclear molecular structures containing R groups (e.g., R-Cl). Another 752 compounds did not contain any carbon (e.g., inorganic ions), and 184 were found to be big molecules, many of them with closed aromatic ring structures that were not accessible for the reaction rules (e.g., fullerene). Sixteen compounds contained only one carbon atom that was not accessible to enzymes (e.g., CFe8S9). The remaining four compounds were found to be chemically synthesized molecules with medical or research applications (Figure 3.3 panel B). Even though these compounds do not seem to have the biochemical capacity to participate in any enzyme-catalyzed reaction, their presence in biological databases can still be justified through their interaction with living organisms.

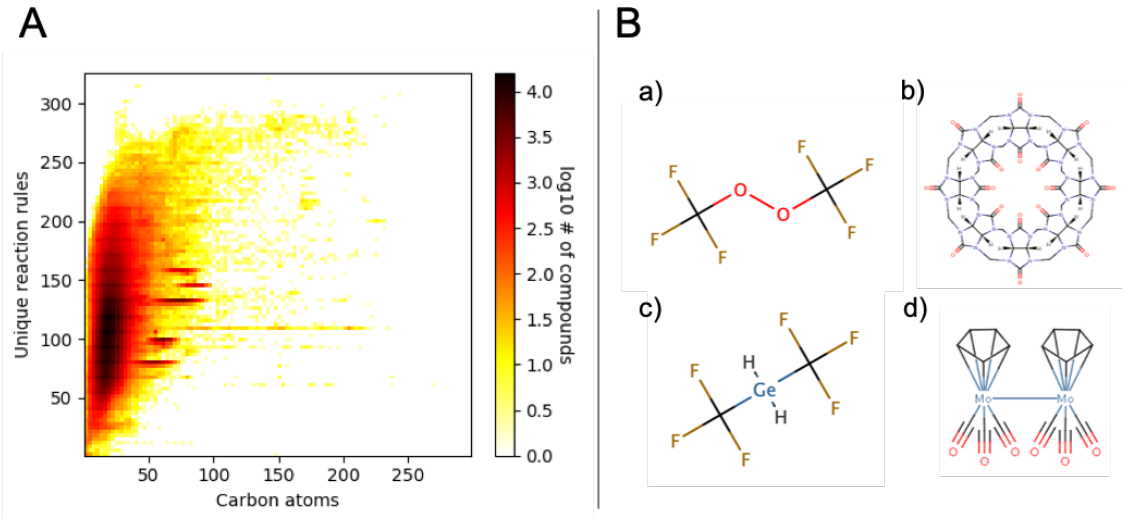


Figure 3.3: Reactive site analysis on bioATLAS compounds (A) Heatmap showing the distribution of compounds as a function of their number of carbon atoms versus the number of reaction rules assigned to them. The color indicates the number of compounds on a logarithmic scale. (B) Four bioactive compounds for which BNICE.ch could not find any reactive site. a) Bis(trifluoromethyl)peroxide(BTP), b) cucurbit[8]uril, c) Bis(trifluoromethyl)germane, d) bis[tricarbonyl(η5-cyclopentadienyl)molybdenum](Mo—Mo).

The number and types of reaction rules assigned to a compound is as an indicator for the diversity of functional groups, or the biochemical versatility, of the molecule. By screening the biological and bioactive molecules for reactive sites, we could show that almost all molecules in bioDB have the potential to undergo biochemical transformations. Using our reactive site screening approach, we can now evaluate their biological reactivity by providing a bioactivity index, which can be calculated as the amount of different reaction rules that can be assigned to a molecule divided by the total number of carbons in the molecule.

Such a bioactivity index could be used in the future to screen big data sets of molecules for potential bioactivity (Appendix, Figure 8.1).

3.4.4 ATLASx predicts 5.3 million novel, hypothetical reactions

In our attempt to map dark matter in metabolism, we used the unified and characterized known biochemical space in bioDB to explore the hypothetical biochemical space by predicting novel, hypothetical reactions from biological and bioactive compounds.

To achieve this, we applied the 490 bidirectional reaction rules from BNICE.ch to the 1,498,307 compounds in bioDB that had at least one reaction rule assigned in the previous step of the workflow. The application of reaction rules produced known and novel, hypothetical reactions leading towards known and novel product molecules. Reactions whose products were part of the biological and bioactive compounds space were stored in the bioATLAS data collection, and reactions whose products that were only part of the chemical compound space were stored in chemATLAS. In total, we reconstructed 11,759 of the metabolic reactions in bioDB and we predicted 5,195,062 novel reactions from biological and bioactive compounds (chemATLAS). Out of these reactions, 1,622,447 (31%) occurred exclusively between biological and bioactive compounds (bioATLAS), and the remaining 3,572,615 reactions involved at least one compound from the chemical space (Table 3.2). 81% of predicted reactions have an estimated Gibbs free energy assigned, and all predicted reactions have a third-level EC number assigned. In terms of compounds, bioATLAS integrates almost two third (844,316 out of 1,500,222) of the compounds fed initially to the workflow. From the remaining 655,906 bioDB compounds, an additional 163,460 were integrated in at least one chemATLAS reaction.

One of the objectives of the ATLAS workflow is to integrate orphan biological and bioactive compounds into the biochemical reaction space. The bioDB counts 1,485,324 orphan compounds that are not involved in any known reaction, even though they are labeled as biological or bioactive molecules. Interestingly, 67% (992,878) out of these orphan compounds could be integrated into at least one novel reaction. We further found that 863,000 compounds, originally only present in PubChem, could be integrated into at least one biochemical reaction, meaning that they are situated only one reaction step away from a known biological or biochemical compound. These compounds are potential candidates for secondary metabolites (e.g., plant natural products), unwanted products of side reactions (i.e., damaged metabolites[40]), or bioactive compounds (e.g., drugs, pesticides) with the capacity to be transformed by enzyme catalysis.

Table 3.2: Compound and reaction statistics for bioDB, bioATLAS and chemATLAS.

	Property	bioDB	bioATLAS	chemATLAS
	Integrated in reaction	14,902	915,372	1,987,019
Compounds	**Total number of compounds**	1,500,222	1,500,222	77,934,143
	Known reactions	56,602	56,602	56,602
	Out of which BNICE.ch curated	*11,759*	*11,759*	*11,759*
Reactions	Novel reactions	0	1,561,139	5,138,460
	Total number of reactions	56,602	1,610,688	5,195,062

3.4.5 Network analysis of the biotransformation network reveals disjoint components

The connectivity of a biochemical reaction network can provide insights into the comprehensiveness our knowledge, and help us to identify missing biochemical links. According to the chemical law of mass conservation, the network representing perfect biochemical knowledge would be fully connected, meaning that every compound (node) is connected to every other compound through a suite of biotransformations (edges).

To create a graph-representation of our reaction database, we employed the concept of atom conservation between substrates and products. This approach that has previously been shown to be relevant in the analysis and search of metabolic networks [41]. We calculated the number of conserved atoms between each possible substrate-product pair in each reaction based on the reaction mechanism encoded in the BNICE.ch reaction rules. For known biological reactions without BNICE.ch reaction mechanism, the number of conserved atoms was estimated by assuming the maximal possible atoms to be conserved. The number of conserved atoms between each substrate-product pair was then used to assign a weight to each edge in the network. The weight, termed Conserved Atom Ratio (CAR), ranges from 0 to 1 and represents the atom conservation between substrate and product. To assess the connectivity of the different networks generated in this study, we excluded edges with a CAR below 0.34, a threshold previously shown to best predict manually curated substrate-product pairs in KEGG[41]. The procedure of network construction is based on the NICEpath methodology [41].

To characterize the different networks scopes in ATLASx, we extracted the network for each of the reaction scopes bioDB, bioATLAS, and chemATLAS (Table 3.3). For each scope, we counted the number of connected components (i.e., disjoint graphs, or islands) in the unweighted networks (Figure 3.4 panel A). We found that

the total number of components increased with the network expansion from bioDB to bioATLAS to chemATLAS. However, the number of components relative to the size of the network, represented by the average number nodes per component, decreased from 23.7 in bioDB to 10.3 in bioATLAS, and increased again to 12.6 in chemATLAS, suggesting that the integration of bioactive compounds created many disconnected islands in the network, which becomes more connected after including chemical compounds. By looking at the size distribution of the components, we found that all three networks were dominated by one big component, followed by a big number of secondary components of maximal 33 compounds involved (Figure 3.4 panel B). While the biggest component in bioDB connected 88% of compounds in the network, this number decreased to 53% in bioATLAS and increased again to 67% in chemATLAS. This result is consistent with the average number of nodes per component, which indicates that integrating bioactives creates a high number of disconnected compound islands, and integrating chemical compounds makes the biochemical network denser by bridging the disconnected islands in bioATLAS. This statement is further confirmed by the diameter metrics: To calculate the diameter of a network, one needs to find all the shortest paths between all the possible combination of nodes in the network. The longest shortest path is called *diameter* of the network, and the average length of shortest paths between any two nodes is called *effective diameter.* Here, we found that the effective diameter is increased in chemATLAS (34 step diameter) compared to bioDB (27 steps) and bioATLAS (27 steps), suggesting expansion of the network towards novel chemistry and integration of previously disconnected components.

Graph theory was used before to analyze the properties of biochemical networks, but these analyses were restricted to either single databases, or performed on specific organisms. Here, we estimated the network properties of known and expanded biochemistry employing state-of-the-art graph-theoretical metrics as well as a robust definition of edges as substrate-product pairs weighted by atom conservation.

Table 3.3: Network statistics of bioDB, bioATLAS and chemATLAS networks.

Network	Property	bioDB	bioATLAS	chemATLAS
Weighted network	Number of nodes	14,902	915,372	1,987,019
	Number of edges (CAR > 0)	62,255	2,624,726	5,849,013
Unweighted network (Only edges with CAR > 0.34)	Number of nodes	14,071	716,924	1,964,445
	Number of edges (CAR > 0.34)	25,597	1,096,283	2,942,679
	Number of components (disjoint graphs)	627	88,487	157,541

Biggest component	Number of nodes	3,359	381,400	1,323,168
	Number of edges	4,405	819,512	2,403,922
	Percent of total number of nodes	88.22%	53.20 %	67.36 %
	Percent of total number of edges	95.82 %	74.75 %	81.69 %
	Diameter (longest shortest path)	27	27	34
	Effective diameter (average shortest path)	7	10	11

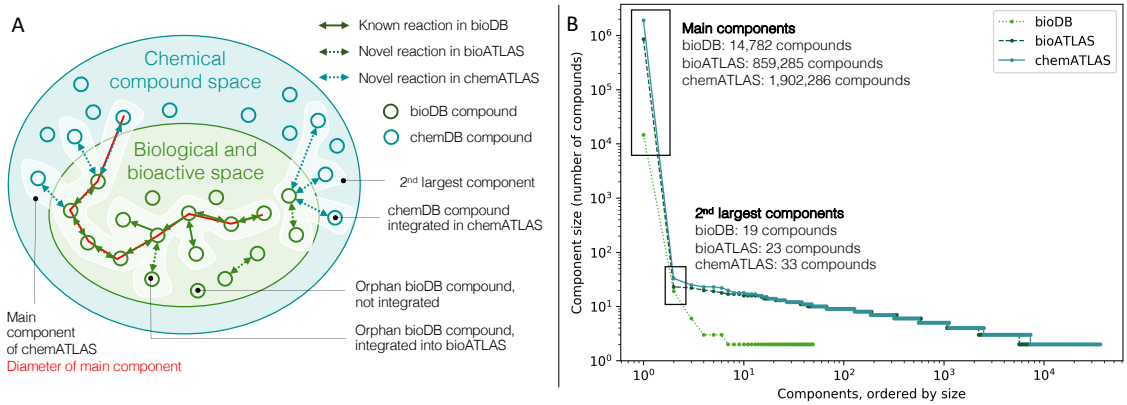


Figure 3.4: Graph-theoretical analysis of biotransformation networks (A) Schematic overview on different statistics and network properties calculated for bioDB, bioATLAS and chemATLAS. Reactions involving one or more chemical compound are assigned to the chemATLAS reaction space (B) Size distribution of disconnected components in the network of each of the three database scopes.

3.4.6 Searching for biological pathways within ATLASx

The quest for novel biosynthesis pathways is crucial for the bioproduction of natural and chemical compounds in chassis organisms, the elucidation of complex natural product biosynthesis, and the study of chemical biodegradation. To search for biological pathways, we used the atom-weighted biochemical transformation networks as established in the previous section for bioDB, bioATLAS, and chemATLAS. Depending on the network scope chosen by the user, the pathways will contain only known reactions (bioDB), include novel reactions between known biological and bioactive compounds (bioATLAS), or even include chemical compounds (chemATLAS).

To benchmark our pathway search, we determined whether ATLASx could recover known pathways from the metabolic pathway database MetaCyc. To do this, 3,149 pathways were collected from MetaCyc, out of which 1,518 matched our curation standards:

- Pathway contains 2 and more reactions (single reaction pathways excluded)
- Pathway does not contain transport reactions and electron-transfer reactions
- Pathway does not contain reactions not indicated in MetaCyc download reactions table (data quality)
- Pathway does not contain the following compound-types as intermediates: compounds with undefined structure, non-carbon compounds, proteins and peptide polymers, RNA molecules, unknown compounds
- Not a circular pathway (excluded as we use loop-less pathway search algorithm and therefore do not target this group)
- Not a polymerization pathway (e.g. bacterial peptidoglycan polymerization)
- Not a light-dependent pathway
- Not a superpathway (pathways consisting of other pathways with no individual unique reaction sequence).

For each pair of precursor-target compounds, we extracted the 100 shortest pathways from ATLASx, and compared them to the original MetaCyc pathway. We were able to both find pathways for 1508 (99%) of precursor-target pairs (Figure 3.5 A) and exactly reconstruct the original MetaCyc pathway for 700 reaction pairs (46%). For another 780 precursor-target pairs (51%), all the required biotransformations were present in the ATLASx network, but the original MetaCyc pathway was not found within the top 100 shortest pathways predicted by ATLASx. The remaining 28 precursor-target pairs (<1%) could only be reconstructed using alternative pathways.

About 80% of reactions in bioDB do not have their reaction mechanism described in BNICE.ch, which could occur for two reasons: (i) the reaction mechanism is unknown or (ii) the reaction mechanism has not yet been added to the BNICE.ch reaction rule collection. Many reactions missing a mechanism are generally also missing further annotation, such as protein sequence and cofactor usage. Depending on our research question, reactions without a confirmed BNICE.ch reaction mechanism can be excluded from pathway searches. By excluding reactions *without* known BNICE.ch mechanisms from the pathway search, we could find pathways that guarantee a biochemical reaction mechanism in each step. By repeating the pathway search for the MetaCyc benchmark set using only bioDB compounds with known BNICE.ch mechanisms, we found connections between precursor and target compounds for 1318 out of 1518 pathways (87%). For 524 MetaCyc pathways (35%), the correct sequence of intermediates were found in the 100 top-ranked pathways. Another 236 pathways (16%) were present in the network as a sequence of biotransformations, but were not found within the top 100 pathways. Finally, for 558 MetaCyc pathways (37%), we only found alternative sequences of intermediates (Figure 3.5 A). We discovered that by excluding reactions without known mechanisms from the network when all required biotransformations were present in the network, we increased the chance of finding the correct pathway within the top 100 pathways. Here, the pathway

search exactly reconstructed 524 out of 760 pathways (69%) for which all the necessary biotransformations were present. In the original network that included all bioDB reactions (with incomplete reaction mechanisms), we only found 700 out of 1480 pathways (47%) that had all biotransformations present. This analysis shows that known biochemical pathways can be diligently reproduced using ATLASx, particularly when reactions without a known mechanism are excluded.

We also investigated how the length of the reference pathway from MetaCyc affects the proportion of reconstructed pathways. We found that even pathways as long as 16 reaction steps could be exactly reconstructed from the original MetaCyc pathway, and that alternative pathways for MetaCyc pathways could be up to 26 reaction steps in length. (Figure 3.5 B). These results show that the performance of the pathway search is not significantly compromised when searching for longer pathways. The pathway search tool is available online at https://lcsb-databases.epfl.ch/Search2. Users can adjust the network scope as discussed above, as well as perform database-specific search scopes for all of the imported reaction databases.

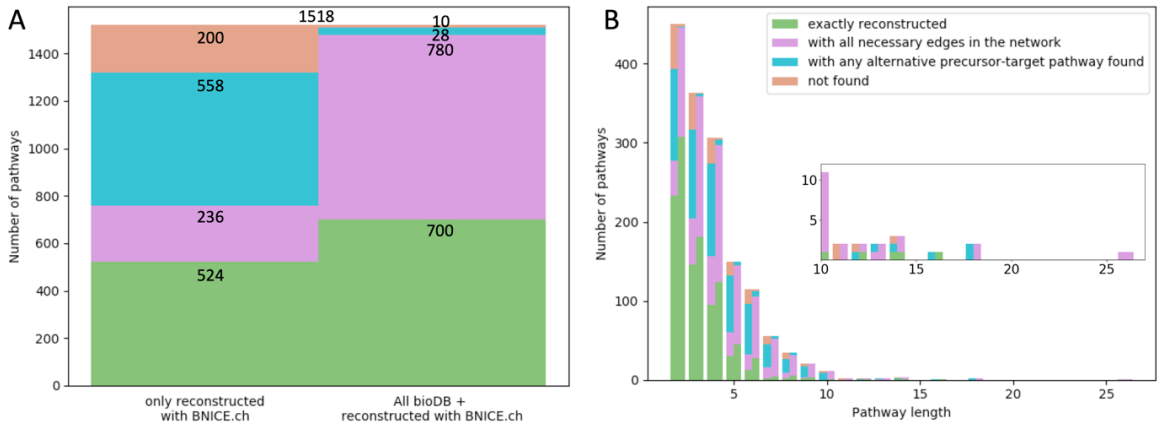


Figure 3.5: Pathway search comparison to dataset of pathways extracted from MetaCyc. (A) Overall statistics for the collected MetaCyc pathways dataset (1518 pathways) coverage with ATLAS pathway search tool (B) Distribution MetaCyc covered pathway depending on the pathway length (only reconstructed with BNICE.ch left side of the bar and all bioDB + reconstructed with BNICE.ch in the right side of the bar as in the overall statistics).

3.4.7 ATLASx fills metabolic gaps and proposes new biosynthesis pathways

ATLASx is a resource with wide-ranging practical applications that include compound classification, metabolic pathway searches, and gap-filling for metabolic models. The following case study presents an example of a practical application that can be performed using the ATLASx web tools.

To illustrate how the unification and expansion of knowledge in ATLASx can inform hypotheses around a given biochemical pathway, we used ATLASx to explore and expand an example of a biosynthetic pathway of interest. We chose the biosynthesis pathway of the anti-fungal and anti-hypertensive compound staurosporine, a secondary metabolite with a complex molecular structure. According to KEGG, the biosynthesis of staurosporine from tryptophan involves 7 reaction steps, however, this biosynthetic pathway is poorly characterized (Figure 3.6 panel B). To explore the biochemical vicinity of this pathway, we retrieved all compounds that were one step away from the original pathway. Out of 861 potential pathway derivatives, 60 were found exclusively through bioDB, 407 compounds were contributed by bioATLAS, and the remaining 394 compounds were only integrated when chemATLAS reactions and compounds were considered (Figure 3.6 panel A). According to our analysis, most derivatives (93%, or 799) were detected around tryptophan. Secondary hubs were found around the precursor K-252c and staurosporine, with each hub contributing 3% (24 compounds) and 3% (22 compounds) to the total number of pathway derivatives, respectively. Intrigued by the high number of potential staurosporine derivatives, we explored four generations of biosynthesis around this molecule (Figure 3.6 panel C) and found 58 derivatives within a distance of four reaction steps. We found 6 staurosporine derivatives within bioDB (4 of them part of the original pathway), 18 derivatives only within bioATLAS, and an additional 34 compounds from chemATLAS. Interestingly, the network exploration converged, and the only derivatives found four steps away from staurosporine were located upstream of the original pathway.

To characterize the potential staurosporine derivatives we identified, we retrieved the number of patents and citations associated with these compounds, which are metrics that have been previously used to assess the “popularity” of compounds. Within the five top-ranked derivatives, we found staurosporine garnered the most attention (29,819 patents and 15,439 citations), followed by 7-hydroxystaurosporine and then K-252c, which is part of the staurosporine synthesis pathway. Midostaurin, a cancer therapeutic and protein kinase inhibitor commercially known as Rydapt, ranked fourth with 158 patents and 570 citations, and was one step away from staurosporine. This analysis illustrates how ATLASx can be used to explore the biochemical vicinity of a compound or pathway and to retrieve relevant information (e.g., citations and patents) from external sources to filter and rank the generated network.

Next, we investigated the capability of ATLASx to detect and bridge knowledge gaps. Out of the 7 reaction steps in the pathway obtained from KEGG, only one reaction is linked to an enzyme. The other 6 reactions are orphan (i.e., no enzyme assigned), out of which 3 reactions are unclear (i.e., no knowledge about reaction mechanism or cofactors involved) (Table 3.4). To show how one can find plausible enzymes(s) for orphan reactions, we examined each orphan reaction within this pathway. First, for reactions with assigned BNICE.ch reaction rules, we applied the computational tool BridgIT to find known reactions with a similar reaction mechanism and structurally similar reactants (see chapter 4). For reaction steps without an assigned

BNICE.ch rule, we searched for pathways that connected reaction intermediates to sequences of known, well-annotated bioDB reactions, or BNICE.ch-reconstructed reactions that provide the basis for robust enzyme prediction with BridgIT.

The first step of the pathway, the conversion of L-tryptophan to IPA imine, is identified with the partial EC number 1.4.3.- by KEGG. The computational tool BridgIT proposed the enzyme 7-chloro-L-tryptophan oxidase (EC 1.4.3.23) as the best candidate to catalyze this first step. This predication was bolstered by a high BridgIT score of 0.95, which indicates that both substrates have a similar reactive site and surrounding structure. While the native function of this proposed enzyme is to convert 7-chloro-L-tryptophan to 2-imino-3-(7-chloroindol-3-yl) propanoate, the activity of this candidate enzyme on L-tryptophan has been proven in a study by Nishizawa *et al.*, suggesting a potential role in this orphan reaction[42]. Another orphan reaction in the staurosporine pathway that was reconstructed by BNICE.ch is the conversion of 3'-demethylstaurosporine to O-demethyl-N-demethyl-staurosporine (step 6). For this reaction, BridgIT suggested that an N-formiminotransferase serves as a catalyzing enzyme (EC 2.1.2.5), although this prediction is accompanied by a relatively low BridgIT score of 0.34. Finally, the last step of the pathway is known to be catalyzed by an O-methyltransferase with EC number 2.1.1.139. In this case, BridgIT successfully mapped this reaction to itself and finds the original EC number. This showcase exemplifies how BridgIT can be used on top of the ATLASx reaction prediction to find enzymes for novel or orphan reactions and to fill gaps in metabolic pathways and networks.

Finally, all of the presented analyses can be performed using the computational tools available online. We provide public access to our database through an online search interface, which includes a powerful pathway search algorithm that can be used for the design of novel metabolic pathways. The web access to ATLASx (https://lcsb-databases.epfl.ch/Atlas2) provides further query tools, such as the ability to identify all reactions associated to a query compound or the ability to extract reactions surrounding a reaction mechanism of interest (i.e., third-level EC number).

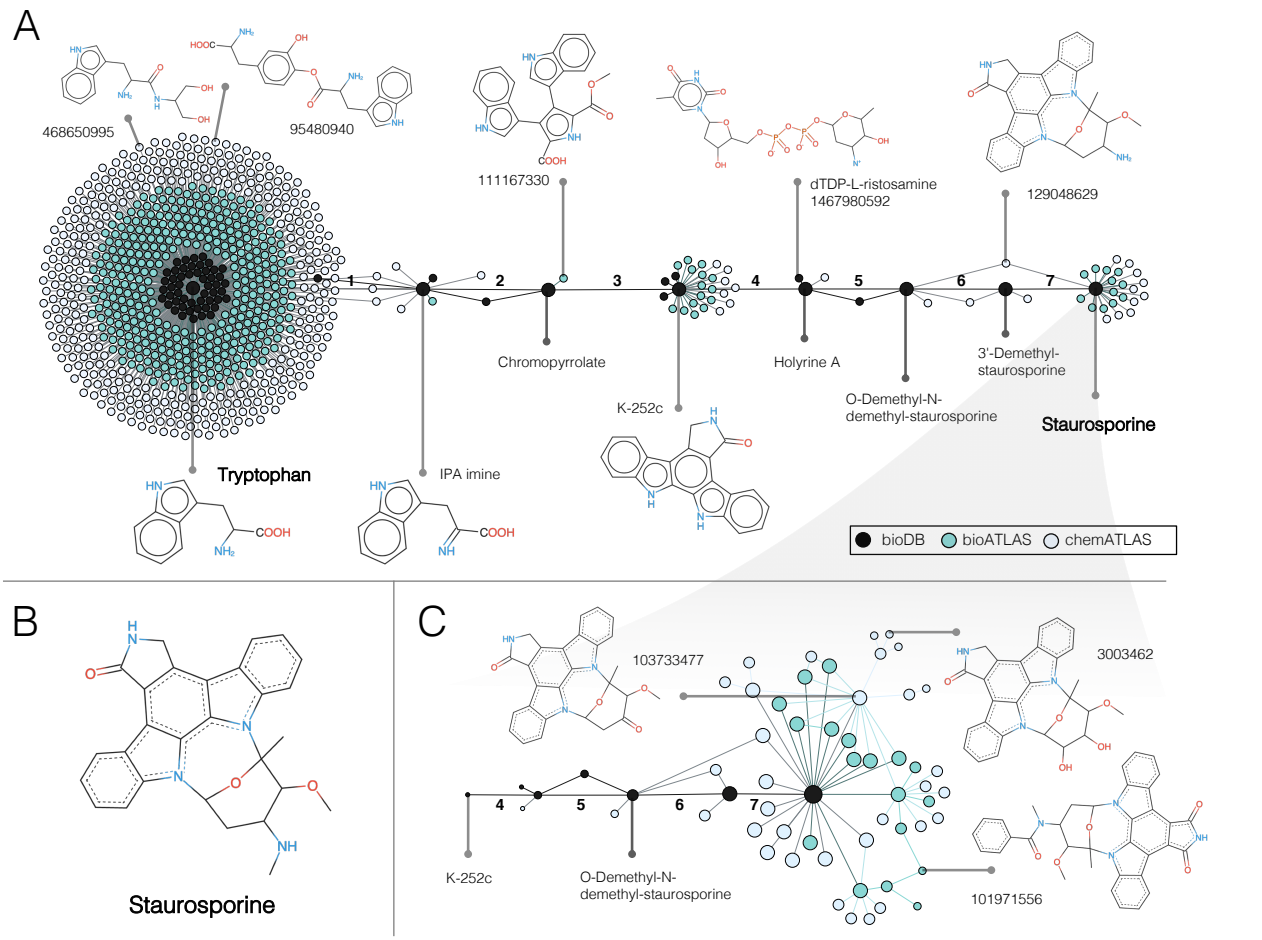


Figure 3.6: Showcase of a pathway expansion for the biosynthesis of the natural product staurosporine. (A) The biosynthesis pathway from tryptophan to staurosporine (obtained from KEGG, steps numbered in bold black) has been expanded for one generation around the native intermediates. (B) The molecular structure of staurosporine. (C) To zoom in specifically on the potential staurosporine biochemistry, the network has been expanded for four generations around the target compound. The size of the nodes representing compounds decreases with each generation.

Table 3.4: Pathway reconstruction and gap-filling within ATLASx for the staurosporine biosynthesis pathway

Step	KEGG ID	EC number	BNICE.ch rule	Top BridgIT hit EC (KEGG ID, score)	Reconstruction within ATLASx
1	R11119	1.4.3.-	1.4.3.-	1.4.3.23 (R09560, 0.95)	Biotransformation with LCSB ID 2600177067
2	R11120				2-step reaction (spontaneous + 1.21.98.2) in bioDB[a]
3	R11121	1.13.12.-			Not reconstructed within ATLASx
4	R11122	2.4.-.-			3-step reaction in chemATLAS[b]
5	R11123				2-step reaction in bioATLAS[c]
6	R11129		2.1.1.-	2.1.2.5 (R03189, 0.34)	Biotransformation with LCSB ID 2600423725
7	R05757	2.1.1.139[43]	2.1.1.-	2.1.1.139 (R05757, 1.00)	Biotransformation with LCSB ID 2600261843

[a] https://lcsb-databases.epfl.ch/Graph2/loadPathway/1/1468050408,1469435049,1468050416/2806125367,2806150968/0
[b] https://lcsb-databases.epfl.ch/Graph2/loadPathway/1/1468050425,1469288899,277921848,1468050433/2603459454,2603467379,2682146339/0
[c] https://lcsb-databases.epfl.ch/Graph2/loadPathway/1/1468050433,1469288674,1468050440/2603455158,2682148818/0

3.5 Conclusion

This work attempts to use biochemical knowledge and biochemical reaction principles to map the hypothetical vicinity of known biochemical databases to address the vast amount of metabolic “dark matter”. Based on 1.5 million known biological and bioactive compounds unified in bioDB, we predicted 1.6 million biochemically possible biotransformations between biological and bioactive compounds using 490 generalized reaction rules (bioATLAS). We then predicted more than 3.6 million reactions that involved compounds from the chemical compound space, resulting in a total of almost 5.2 million in chemATLAS. From this new wealth of information, we extracted insightful numbers on the reactivity and connectivity of biologically relevant molecules.

Assessing the composition of metabolic “dark matter” is by definition difficult, since we lack a way to quantify the unknowns a priori. Fortunately, biochemical data collected and generated from our database allows us to answer a broad range of questions regarding the biochemical reactivity of compounds, the expansion of biochemical space from a graph-theoretical perspective, and the characteristics of our hypothetical reaction network. Potential applications of ATLASx include the prediction of bioproduction or biodegradation pathways involved in the transformation of commodity and specialty chemicals, pharmaceuticals, and plastics. ATLASx can also be used to discover the biosynthesis routes of poorly characterized secondary metabolites, and systematically fill in knowledge gaps surrounding metabolic models. Using ATLASx, one can expand the network around all compounds within a given metabolic model, remove dead-end metabolites, and then examine the new, expanded model for potential shortcuts, enhanced predictions, and enzymatic promiscuity.

Since we successfully integrated tens of thousands of chemical compounds into a biochemical network, we hypothesize that many compounds are not yet part of any database, even though they potentially exist in nature or could be created by metabolic engineering. While the integration and accurate prediction of hypothetical compound structures remains an open challenge, ATLASx provides the necessary tools and conceptual framework to predict hypothetical compounds reliably in the future. In order to properly meet that future, ATLASx is designed as a dynamic database, and can be continuously expanded around biochemical pathways or compound classes of interest. We believe that predictive biochemistry is crucial for the advancement of synthetic biology and metabolic engineering, and hope that ATLASx can provide reliable reaction and pathway predictions for the scientific community.

3.6 References

[1] N. Hadadi, J. Hafner, A. Shajkofci, A. Zisaki, and V. Hatzimanikatis, “ATLAS of Biochemistry: A repository of all possible biochemical reactions for synthetic biology and etmabolic engineering studies,” *ACS Synth. Biol.*, vol. 5, no. 10, pp. 1155–1166, Oct. 2016, doi: 10.1021/acssynbio.6b00054.

[2] J. Hafner, H. MohammadiPeyhani, A. Sveshnikova, A. Scheidegger, and V. Hatzimanikatis, “Updated ATLAS of Biochemistry with New Metabolites and Improved Enzyme Prediction Power,” *ACS Synth. Biol.*, vol. 9, no. 6, pp. 1479–1482, 06-19 2020, doi: 10.1021/acssynbio.0c00052.

[3] J. L. Ross, “The Dark Matter of Biology,” *Biophys. J.*, vol. 111, no. 5, pp. 909–916, Sep. 2016, doi: 10.1016/J.BPJ.2016.07.037.

[4] P. Kapranov and G. St. Laurent, “Dark Matter RNA: Existence, Function, and Controversy,” *Front. Genet.*, vol. 3, p. 60, Apr. 2012, doi: 10.3389/fgene.2012.00060.

[5] L. Solden, K. Lloyd, and K. Wrighton, “The bright side of microbial dark matter: lessons learned from the uncultivated majority,” *Curr. Opin. Microbiol.*, vol. 31, pp. 217–226, Jun. 2016, doi: 10.1016/J.MIB.2016.04.020.

[6] R. A. Notebaart, B. Kintses, A. M. Feist, and B. Papp, “Underground metabolism: network-level perspective and biotechnological potential,” *Curr. Opin. Biotechnol.*, vol. 49, pp. 108–114, Feb. 2018, doi: 10.1016/j.copbio.2017.07.015.

[7] J. Rosenberg and F. M. Commichau, “Harnessing Underground Metabolism for Pathway Development,” *Trends Biotechnol.*, vol. 37, no. 1, pp. 29–37, Jan. 2019, doi: 10.1016/J.TIBTECH.2018.08.001.

[8] C. Lerma-Ortiz *et al.*, “‘Nothing of chemistry disappears in biology’: the Top 30 damage-prone endogenous metabolites,” *Biochem. Soc. Trans.*, vol. 44, no. 3, pp. 961–971, Jun. 2016, doi: 10.1042/BST20160073.

[9] N. E. Lewis and A. M. Abdel-Haleem, “The evolution of genome-scale models of cancer metabolism,” *Front. Physiol.*, vol. 4, p. 237, Sep. 2013, doi: 10.3389/fphys.2013.00237.

[10] R. R. Stanway *et al.*, “Genome-Scale Identification of Essential Metabolic Processes for Targeting the Plasmodium Liver Stage,” *Cell*, vol. 179, no. 5, pp. 1112-1128.e26, Nov. 2019, doi: 10.1016/j.cell.2019.10.030.

[11] A. G. Atanasov *et al.*, “Discovery and resupply of pharmacologically active plant-derived natural products: A review,” *Biotechnol. Adv.*, vol. 33, no. 8, pp. 1582–1614, Dec. 2015, doi: 10.1016/J.BIOTECHADV.2015.08.001.

[12] S. G. Oliver, “From DNA sequence to biological function,” *Nature*, vol. 379, no. 6566, pp. 597–600, Feb. 1996, doi: 10.1038/379597a0.

[13] M. Y. Galperin and E. V. Koonin, “From complete genome sequence to ‘complete’ understanding?,” *Trends Biotechnol.*, vol. 28, no. 8, pp. 398–406, Aug. 2010, doi: 10.1016/J.TIBTECH.2010.05.006.

[14] I. M. Keseler *et al.*, “The EcoCyc database: reflecting new knowledge about *Escherichia coli* K-12,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D543–D550, Jan. 2017, doi: 10.1093/nar/gkw1003.

[15] B. O. Bachmann, “Biosynthesis: Is it time to go retro?,” *Nat. Chem. Biol.*, vol. 6, no. 6, pp. 390–393, Jun. 2010, doi: 10.1038/nchembio.377.

[16] N. Hadadi and V. Hatzimanikatis, “Design of computational retrobiosynthesis tools for the design of de novo synthetic pathways,” *Curr. Opin. Chem. Biol.*, vol. 28, pp. 99–104, Oct. 2015, doi: 10.1016/J.CBPA.2015.06.025.

[17] L. Wang, C. Y. Ng, S. Dash, and C. D. Maranas, “Exploring the combinatorial space of complete pathways to chemicals.,” *Biochem. Soc. Trans.*, vol. 46, no. 3, pp. 513–522, Jun. 2018, doi: 10.1042/BST20170272.

[18] G.-M. M. Lin, R. Warden-Rothman, and C. A. Voigt, “Retrosynthetic design of metabolic pathways to chemicals not found in nature,” *Curr. Opin. Syst. Biol.*, vol. 14, pp. 82–107, Apr. 2019, doi: 10.1016/J.COISB.2019.04.004.

[19] J. G. Jeffryes, S. M. D. Seaver, J. P. Faria, and C. S. Henry, “A pathway for every product? Tools to discover and design plant metabolism,” *Plant Sci.*, Mar. 2018, doi: 10.1016/J.PLANTSCI.2018.03.025.

[20] V. Hatzimanikatis, C. Li, J. A. Ionita, C. S. Henry, M. D. Jankowski, and L. J. Broadbelt, “Exploring the diversity of complex metabolic networks,” *Bioinformatics*, vol. 21, no. 8, pp. 1603–1609, Apr. 2005, doi: 10.1093/bioinformatics/bti213.

[21] M. Tokić *et al.*, “Discovery and evaluation of biosynthetic pathways for the production of five methyl ethyl ketone precursors,” *ACS Synth. Biol.*, p. acssynbio.8b00049, Jul. 2018, doi: 10.1021/acssynbio.8b00049.

[22] M. A. Campodonico, B. A. Andrews, J. A. Asenjo, B. O. Palsson, and A. M. Feist, “Generation of an atlas for commodity chemical production in Escherichia coli and a novel pathway prediction algorithm, GEM-Path,” *Metab. Eng.*, vol. 25, pp. 140–158, Sep. 2014, doi: 10.1016/J.YMBEN.2014.07.009.

[23] S. Ding *et al.*, “novoPathFinder: a webserver of designing novel-pathway with integrating GEM-model,” *Nucleic Acids Res.*, no. 1, 2020, doi: 10.1093/nar/gkaa230.

[24] A. Kumar, L. Wang, C. Y. Ng, and C. D. Maranas, “Pathway design using de novo steps through uncharted biochemical spaces,” *Nat. Commun.*, vol. 9, no. 1, p. 184, Jan. 2018, doi: 10.1038/s41467-017-02362-x.

[25] T. V. Sivakumar, V. Giri, J. H. Park, T. Y. Kim, and A. Bhaduri, “ReactPRED: A tool to predict and analyze biochemical reactions,” *Bioinformatics*, p. btw491, Aug. 2016, doi: 10.1093/bioinformatics/btw491.

[26] B. Delépine, T. Duigou, P. Carbonell, and J.-L. Faulon, “RetroPath2.0: A retrosynthesis workflow for metabolic engineers,” *Metab. Eng.*, vol. 45, pp. 158–170, Jan. 2018, doi: 10.1016/j.ymben.2017.12.002.

[27] M. Koch, T. Duigou, and J.-L. Faulon, “Reinforcement Learning for Bio-Retrosynthesis,” *bioRxiv*, p. 800474, Nov. 2019, doi: 10.1101/800474.

[28] J. D. Tyzack, A. J. M. Ribeiro, N. Borkakoti, and J. M. Thornton, “Exploring Chemical Biosynthetic Design Space with Transform-MinER,” *ACS Synth. Biol.*, vol. 8, no. 11, pp. 2494–2506, Nov. 2019, doi: 10.1021/acssynbio.9b00105.

[29] J. Wicker *et al.*, “enviPath – The environmental contaminant biotransformation pathway resource,” *Nucleic Acids Res.*, p. gkv1229, Nov. 2015, doi: 10.1093/nar/gkv1229.

[30] J. G. Jeffryes *et al.*, “MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics,” *J. Cheminformatics*, vol. 7, no. 1, p. 44, Dec. 2015, doi: 10.1186/s13321-015-0087-1.

[31] V. Hatzimanikatis, C. Li, J. A. Ionita, C. S. Henry, M. D. Jankowski, and L. J. Broadbelt, “Exploring the diversity of complex metabolic networks,” *Bioinformatics*, vol. 21, no. 8, pp. 1603–1609, Apr. 2005, doi: 10.1093/bioinformatics/bti213.

[32] S. D. Finley, L. J. Broadbelt, and V. Hatzimanikatis, “Computational framework for predictive biodegradation,” *Biotechnol. Bioeng.*, vol. 104, no. 6, pp. 1086–1097, 2009, doi: 10.1002/bit.22489.

[33] C. S. Henry, L. J. Broadbelt, and V. Hatzimanikatis, “Discovery and analysis of novel metabolic pathways for the biosynthesis of industrial chemicals: 3-hydroxypropanoate,” *Biotechnol. Bioeng.*, vol. 106, no. 3, pp. 462–473, 2010, doi: 10.1002/bit.22673.

[34] K. C. Soh and V. Hatzimanikatis, “DREAMS of metabolism,” *Trends Biotechnol.*, vol. 28, no. 10, pp. 501–508, Oct. 2010, doi: 10.1016/j.tibtech.2010.07.002.

[35] M. Tokić *et al.*, “Discovery and evaluation of biosynthetic pathways for the production of five methyl ethyl ketone precursors,” *ACS Synth. Biol.*, vol. 7, no. 8, pp. 1859–1873, 2018, doi: 10.1021/acssynbio.8b00049.

[36] M. Kanehisa and S. Goto, “KEGG: kyoto encyclopedia of genes and genomes,” *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, Jan. 2000.

[37] K. R. Choi, W. D. Jang, D. Yang, J. S. Cho, D. Park, and S. Y. Lee, “Systems metabolic engineering strategies: integrating systems and synthetic biology with metabolic engineering,” *Trends Biotechnol.*, vol. 37, no. 8, pp. 817–837, Feb. 2019, doi: 10.1016/J.TIBTECH.2019.01.003.

[38] S. Y. Lee *et al.*, “A comprehensive metabolic map for production of bio-based chemicals,” *Nat. Catal.*, vol. 2, no. 1, pp. 18–33, Jan. 2019, doi: 10.1038/s41929-018-0212-4.

[39] X. Yang *et al.*, “Systematic design and in vitro validation of novel one-carbon assimilation pathways,” *Metab. Eng.*, vol. 56, pp. 142–153, Dec. 2019, doi: 10.1016/J.YMBEN.2019.09.001.

[40] C. L. Linster, E. Van Schaftingen, and A. D. Hanson, “Metabolite damage and its repair or pre-emption,” *Nature Chemical Biology*, vol. 9, no. 2. Nature Publishing Group, pp. 72–80, Feb. 18, 2013, doi: 10.1038/nchembio.1141.

[41] Jasmin Hafner, “MODELING, PREDICTING AND MINING METABOLISM AT ATOM-LEVEL RESOLUTION,” EPFL, 2020.

[42] T. Nishizawa, C. C. Aldrich, and D. H. Sherman, “Molecular Analysis of the Rebeccamycin l-Amino Acid Oxidase from Lechevalieria aerocolonigenes ATCC 39243,” *J. Bacteriol.*, vol. 187, no. 6, pp. 2084–2092, Mar. 2005, doi: 10.1128/JB.187.6.2084-2092.2005.

[43] S. WEIDNER, M. KITTELMANN, K. GOEKE, O. GHISALBA, and H. ZÄHNER, “3’-Demethoxy-3’-hydroxystaurosporine-O-methyltransferase from Streptomyces longisporoflavus Catalyzing the Last Step in the Biosynthesis of Staurosporine.,” *J. Antibiot. (Tokyo)*, vol. 51, no. 7, pp. 679–682, Jul. 1998, doi: 10.7164/antibiotics.51.679.

Chapter 4 BridgIT: Enzyme annotation for orphan and novel reactions using knowledge of substrate reactive sites

Orphan reactions find a home

The following chapter is published in the journal The Proceedings of the National Academy of Sciences (PNAS). Dr. Noushin Hadadi contributed in conceptualization and planning of this project. Dr. Ljubisa Miskovic provided the results of BLAST analysis, and Dr. Marianne Seijo implemented the standard fingerprint and contributed in the code development. The author of this thesis has been in charge of pipeline development, implementation of reactive site centric fingerprints and data analysis. This project was equally led by Dr. Noushin Hadadi and the author. Prof. Vassily Hatzimanikatis supervised the project as well as the completion of the manuscript. The tool developed here, named BridgIT, is currently hosted on the code sharing platform c4science.

Full list of authors in this paper: N. Hadadi†, H. MohammadiPeyhani†, L. Miskovic, M. Seijo, and V. Hatzimanikatis, "Enzyme annotation for orphan and novel reactions using knowledge of substrate reactive sites," Proc. Natl. Acad. Sci., p. 201818877, Mar. 2019, doi: 10.1073/pnas.1818877116 († contributed equally, * corresponding author).*

4.1 Introduction

Recent advances in synthetic biochemistry have resulted in a wealth of novel hypothetical enzymatic reactions that are not matched to protein-encoding genes, deeming them "orphan". A large number of known metabolic reactions are also orphan, leaving important gaps in metabolic network maps. Proposing genes for the catalysis of orphan reactions is critical for applications ranging from biotechnology to medicine. This chapter starts with an introduction on the enzymatic gaps in biochemistry, followed by short review on recent advances in the field of enzyme prediction. Next, we introduce a novel computational method, BridgIT, to identify potential enzymes of orphan reactions and nearly all theoretically possible biochemical transformations, and we validate predications of BridgIT within several large-scale analyses.

4.1.1 Catalytic dark matter

Genome-scale reconstructions of metabolic networks can be used to correlate the genome with the observed physiology, though this hinges on the completeness and accuracy of the sequenced genome annotations.

Orphan reactions, which are enzymatic reactions without protein sequences or genes associated with their functionality, are common and can be found in the genome-scale reconstructions of even well-characterized organisms, such as *Escherichia coli* [1]. Recent publications reported that 40-50% of the enzymatic reactions cataloged in the Kyoto Encyclopedia of Genes and Genomes (KEGG) [2] lack an associated protein sequence [3][4].

Problems with orphan-like reactions can also arise in areas such as bioremediation, synthetic biology, and drug discovery, where exploring the potential of biological organisms beyond their natural capabilities has prompted the development of tools that can generate *de novo* hypothetical enzymatic reactions and pathways [5]–[11], [12, p. 4], [13]–[15]. These *de novo* reactions are behind many success stories in biotechnology, and they can also be used in the gap-filling of metabolic networks [6], [12], [13], [15]–[18]. While these enzymatic reactions have well-explained biochemistry that can conceivably occur in metabolism, they are essentially orphan reactions because they have no assigned enzyme or corresponding gene sequence. The lack of protein-encoding genes associated with the functionality of these *de novo* reactions limits their applicability for metabolic engineering, synthetic biology applications, and the gap-filling of genome scale models [19]. A method for associating *de novo* reactions to similarly occurring natural enzymatic reactions would allow for the direct experimental implementation of the discovered novel reactions or assist in designing new proteins capable of catalyzing the proposed biotransformation.

4.1.2 Computational approaches

Computational methods for identifying candidate genes of orphan reactions have mostly been developed based on protein sequence similarity [3], [20]–[22]. The two predominant classes of these sequence-based methods revolve around gene/genome analysis [22]–[25] and metabolic information [26], [27]. Several bioinformatics methods combine different aspects of these two classes, such as gene clustering, gene co-expression, phylogenetic profiles, protein interaction data, and gene proximity, for assigning genes and protein sequences to orphan reactions [28]–[31]. All of these methods use the concept of *sequence similarity*. Within this concept, homology between two sequences, one orphan and one well-characterized, is inferred when the two share more similarity than it would be expected by chance [32]. Then, the biochemical function is assigned to the orphan protein sequence assuming that homologous sequences have similar functions. This can be problematic because many known enzymatic activities are still missing an associated gene due to annotation errors, the incompleteness of gene sequences [33], and the fact that homology-based methods cannot annotate orphan protein sequences with no or little sequence similarity to known enzymes [3], [34]. Moreover, sequence similarity methods can provide inaccurate results, as small changes in key residues might greatly alter enzyme functionality [35], and also it is a common observation that vastly different protein sequences can exhibit the same fold and, therefore, have similar catalytic activity even though they look very different [36], [37]. In addition, these methods are not suitable for the annotation of *de novo* reactions since

current pathway prediction tools only provide information about enzyme catalytic biotransformations and not about their sequences.

These shortcomings motivated the development of alternative computational methods based on the *structural similarity of reactants and products* for identifying candidate protein sequences for orphan enzymatic reactions [31], [35], [38]–[42]. The idea behind these approaches was to assess the similarity of two enzymatic reactions via the similarity of their reaction fingerprints, i.e., the mathematical descriptors of the structural and topological properties of the participating metabolites [43], which could eliminate the problems associated with non-matching or unassigned protein sequences. In such methods, the reaction fingerprint of an orphan reaction is compared with a set of non-orphan reference-reaction fingerprints, and the genes of the most similar reference reactions are then assigned as promising candidate genes for the orphan reaction. Reaction fingerprints can be generated based on different similarity metrics, such as the bond change, reaction center, or structural similarity [42].

One class of reaction-fingerprint computational methods compares all of the compounds participating in reactions [42], which includes both reactants and cofactors. The application of this group of methods is restricted to specific enzymatic reactions that do not involve large cofactors [31], [35], [38]–[42]. This is because the structural information of the large cofactors overwhelmingly contributes to the corresponding reconstructed reaction-fingerprint, and consequently, reactions with similar cofactors will inaccurately be classified as similar (35–38).

Another class of reaction-fingerprint methods uses the chemical structures of reactant pairs for comparison [40]. While these methods can be applied to all classes of enzymatic reactions, they neglect the crucial role of cofactors in the reaction mechanism. Moreover, neither of these two classes of methods have been employed for assigning protein sequences to *de novo* reactions [40].

4.1.3 BridgIT

In this chapter, we introduce a novel computational method, BridgIT, that links orphan reactions and *de novo* reactions, predicted by pathway design tools such as BNICE.ch [16], Retropath2 [15], DESHARKY [10], and SimPheny [12], with well-characterized enzymatic reactions and their associated genes. BridgIT uses reaction fingerprints to compare enzymatic reactions and is inspired by the “lock and key” principle that is used in protein docking methods [44] wherein the enzyme binding pocket is the “lock” and the ligand is a “key”. If a molecule has the same reactive sites and a similar surrounding structure as the native substrate of a given enzyme, it is then rational to expect that the enzyme will catalyze the same biotransformation on this molecule. Following this reasoning, BridgIT uses the structural similarity of the reactive sites of participating substrates together with their surrounding structure as a metric for assessing the similarity of enzymatic reactions. It is substrate-reactive-site centric, and its reaction fingerprints reflect the specificities of

biochemical reaction mechanisms that arise from the type of enzymes catalyzing those reactions. BridgIT introduces an additional level of specificity into reaction fingerprints by capturing critical information about the enzyme binding pocket. More precisely, BridgIT allows us to capture approximately the 2D structure of the enzyme binding pocket by incorporating the information about sequences of atoms and bonds around the substrate reactive site.

Through several studies, we demonstrated the effectiveness of utilizing the BridgIT fingerprints for mapping novel and orphan reactions to the known biochemistry. These reactions are mapped according to the enzyme commission (EC) [45] number, which is an existing numerical classification scheme for enzyme-based reactions. The EC number can classify enzymes at up to four levels, with a one-level classification being the most general and a four-level classification being the most specific, and these enzyme-based reactions are then represented by four numbers, one for each level, separated by periods (e.g. 1.1.1.1). We show that BridgIT is capable of correctly predicting enzymes with an identical third-level EC number, indicating a nearly identical type of enzymatic reaction, for 94% of orphan reactions from KEGG 2011 that became non-orphan in KEGG 2016. This result validates the consistency of the sequences predicted by BridgIT with the experimental observations, and it further suggests that BridgIT can provide enzyme sequences for catalyzing nearly all orphan reactions. We also studied how the size of the BridgIT fingerprint impacts the BridgIT predictions. We show that BridgIT correctly identifies protein sequences using fingerprints that describe the neighborhood up to six bonds away from the atoms of the reactive site. Strikingly, we also find that it is sufficient to use the information of only three bonds around the atoms of the reactive sites of substrates to accurately identify protein sequences for 93% of the analyzed reactions.

Finally, to indicate the power of this computational technique, we applied BridgIT to the study of all of the 137,000 novel reactions from the ATLAS of biochemistry, a database of all theoretically possible biochemical reactions [46], most of which have no current route to their synthesis or development. Using our technology, we provide candidate enzymes that can potentially catalyze the biotransformation of these reactions to the research community, which should provide a basis for the engineering and development of novel enzyme-catalyzed biotransformations.

4.2 Materials and Methods

4.2.1 BridgIT workflow

The BridgIT workflow together with an example of its application on an orphan reaction is demonstrated in Figure 4.1. BridgIT is organized into four main steps:

1. reactive site identification,
2. reaction fingerprint construction,

3. reaction similarity evaluation,
4. scoring, ranking, and gene assignment.

The inputs of the workflow are (i) an orphan or a novel reaction and (ii) the collection of BNICE.ch generalized enzyme reaction rules. As explained in chapter 2, these reaction rules assemble biochemical knowledge distilled from the biochemical reaction databases, and they are used to discover *de novo* enzymatic reactions as well as predict all possible pathways from known compounds to target molecules [16], [46], [47]. Here, we used the generalized enzyme reaction rules to extract information about the reactive sites of substrates participating in an orphan or a novel reaction, and we integrated it into the BridgIT reaction fingerprints (Figure 4.1, panels 1 and 2). We then compared the obtained BridgIT reaction fingerprints to the ones from the reference reaction database based on the Tanimoto similarity scores (Figure 4.1, panel 3). A Tanimoto score near 0 designates reactions with no or low similarity, whereas a score near 1 designates reactions with high similarity. We used these scores to rank the assigned reactions from the reference reaction database, and we identified the enzymes associated with the highest-ranked reference reactions as candidates for catalyzing the analyzed orphan or novel reaction (Figure 4.1, panel 4). In the next sections, we discuss the reconstructions and testing of the various components of BridgIT as well as the results of our main analyses. A web-tool of BridgIT can be consulted at [http://lcsb-databases.epfl.ch/pathways/Bridgit.](http://lcsb-databases.epfl.ch/pathways/Bridgit)

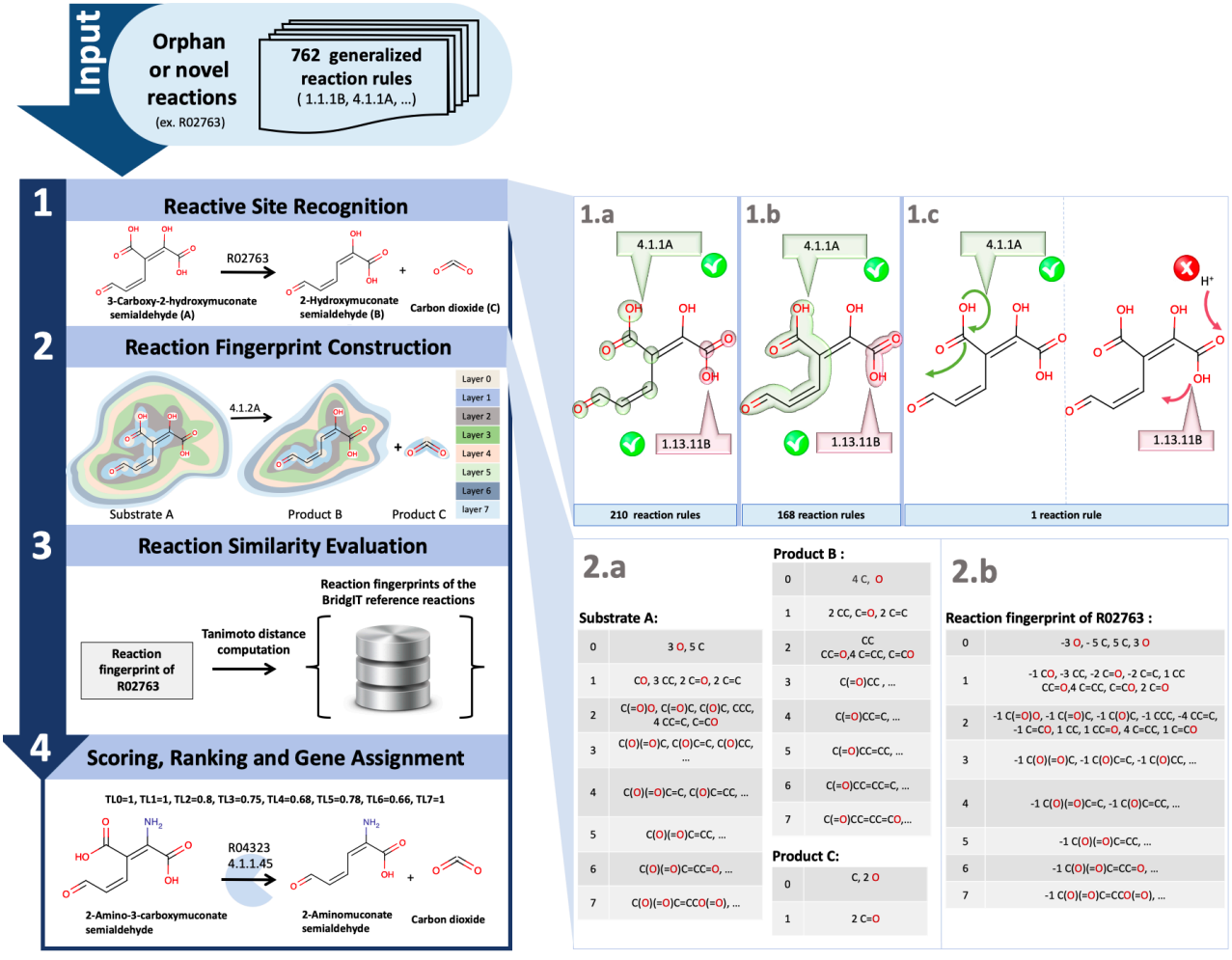


Figure 4.1: Main steps of the BridgIT workflow: (1) reactive site recognition for an input reaction (*de novo* or orphan); (2) reaction fingerprint construction; (3) reaction similarity evaluation; and (4) sorting, ranking and gene assignment. Panels 1.a to 1.c illustrate the procedure of the identification of reactive sites for the orphan reaction R02763. Panel 1.a: Two candidate reactive sites of 3-Carboxy-2-hydroxymuconate semialdehyde (substrate A) that were recognized by the rules 4.1.1. (green) and 1.13.11 (red). Panel 1.b: Both rules recognized the connectivity of atoms within two candidate reactive sites. Panel 1.c: Only reaction rule 4.1.1. can explain the transformation of substrate A to products. Panel 2.a shows the fragmentation of reaction compounds, whereas panel 2.b illustrates the mathematical representations of the corresponding BridgIT reaction fingerprints.

4.2.1.1 Reactive site identification

An enzymatic reaction occurs when its substrate(s) fits into the binding site of an enzyme. Since the structure and geometry of the binding sites of enzymes are complex and most of the time not fully characterized, we proposed focusing on the similarity of the reactive sites of their substrates. Following this, we used the expert-curated, generalized reaction rules of BNICE.ch to identify the reactive sites of substrates. These reaction rules have third-level EC identifiers, e.g., EC 1.1.1, and they encompass the following biochemical knowledge of enzymatic reactions: (i) the information about atoms of the substrate's reactive site; (ii) their connectivity (atom-bond-atom); and (iii) the exact information of bond breakage and formation during the reaction.

Given a novel or orphan reaction, the reactive sites of its substrate(s) are identified in three steps. In the first step, the BNICE.ch generalized reaction rules that can be applied to groups of atoms from the analyzed substrates are identified. Then, the information about the identified rules and the corresponding groups of atoms is stored. Subsequently, these groups of atoms are then referred to as the candidate substrate reactive sites. In the second step, among the identified rules, only the ones that can recognize the connectivity between the atoms of the candidate substrate reactive sites are kept. In the third step, whether the biotransformation of a substrate(s) to a product(s) can be explained by the rules retained after the second step is tested. The candidate reactive sites corresponding to the rules that have passed the three-step test are validated and used for the construction of reaction fingerprints.

We illustrate this procedure on an orphan reaction R02763, which catalyzes the conversion of 3-Carboxy-2-hydroxymuconate semialdehyde (substrate A) to 2-Hydroxymuconate semialdehyde and carbon dioxide (Figure 4.1). In the first step, 210 rules were identified that could be applied to groups of atoms of substrate A (Figure 4.1, panel 1a). Out of the 210 rules, 168 matched the connectivity (Figure 4.1, panel 1b). Finally, the 168 reaction rules were applied to substrate A for bond breaking and formation comparisons, and one rule could explain the transformation of substrate A to the products (Figure 4.1, panel 1c).

4.2.1.2 Reaction fingerprint construction

After recognition of the reactive site, the information of atom-bond configuration inside and also in neighbourhood of the reactive site are stored in a mathematical description, called fingerprint.

Molecular fingerprints, which are the linear representations of the structures of molecules, have been used in many methods and for different applications, especially for structural comparison of compounds [48], [49]. One of the most commonly used molecular fingerprints is the Daylight fingerprint [48], and it decomposes a molecule into eight layers starting from layer zero that accounts only for atoms. Layer 1 expands one bond away from all of the atoms and accounts for atom-bond-atom connections. This procedure is continued until layer 7, which includes seven connected bonds from each atom. There are two types of Daylight reaction fingerprints: (i) structural reaction fingerprints, which are simple combinations of reactant and product fingerprints, and (ii) reaction difference fingerprints, which are the algebraic summation of reactant and product fingerprints multiplied by their stoichiometry coefficients in the reaction. In this study, we propose a modified version of the reaction difference fingerprint. The procedure for formulating BridgIT reaction fingerprints is demonstrated through an example reaction (Figure 4.1, panel 2).

Starting from the atoms of the identified substrate reactive site, eight description layers of the molecule were formed, where different layers consisted of fragments with different lengths. Fragments were composed of atoms connected through unbranched sequences of bonds. Depending on the number of bonds included in the fragments, different description layers of a molecule were formed as follows:

Layer 0: Describes the type of each atom of the reactive site together with its count. For example, the substrate of the example reaction at layer 0 was described as 3 oxygens and 5 carbon atoms (Figure 4.1, panel 2a).

Layer 1: Describes the type and count of each bond between pairs of atoms in the reactive site. In the example, the substrate at layer 1 was described with six fragments of length 1: 1 C-O, 3 C-C, 2 C=O and 1 C=C bond (Figure 4.1, panel 2a). Fragments are shown by their SMILES molecular representation [50]. In order to convert SMILES to canonical SMILES we used Open Babel C++ library [49].

Layer 2: Describes the type and count of fragments with three connected atoms. While layers 0 and 1 described the atoms of reactive sites, starting from layer 2, atoms that were outside of the reactive site were also described. In the illustrated example, there were six different fragments of this type (Figure 4.1, panel 2a).

The same procedure was used to describe the molecules up to layer 7. Interestingly, and consistent with the previously reported result [43], we found that the 7-layer description was good enough to capture the structure of most of the metabolites in biochemical reactions, therefore providing a precise reaction

fingerprint (discussed in section 4.3.3) . Note that not all description layers are needed to describe less complex molecules. For example, product C (carbon dioxide) was fully described using only layer 0 and layer 1 (Figure 4.1, panel 2a). For very large molecules, the description layers that contain fragments with more than 8 connected atoms can be used.

For each layer, the substrate set was formed by merging all of the fragments, their type, and their count in the substrate molecules of the reaction, and the product set was formed by merging all of the fragments (type and count) in the product molecules of the reaction. In both sets, the count of each fragment was multiplied by the stoichiometric coefficients of the corresponding compound in the reaction. Finally, the reaction fingerprints were created by summing the fragments of the substrate and product sets for each layer (Figure 4.1, panel 2b).

Introducing the specificity of reactive sites into the reaction fingerprint allows BridgIT to capitalize on the information about enzyme binding pockets [16]. To keep this valuable information throughout the generation of reaction fingerprints, BridgIT does not consider the atoms of the reactive site(s) when performing the algebraic summation of the substrate and product set fragments. Consequently, the BridgIT algorithm enables retaining, tracking, and emphasizing the information of the reactive site(s) in all of the layers of the reaction fingerprint, which distinguishes it from the existing methods.

4.2.1.3 Reaction similarity evaluation

The similarity of two reactions was quantified using the similarity score between their fingerprints, subsequently referred to as reaction fingerprints A and B. In this study, the Tanimoto score, which is an extended version of the Jaccard coefficient and cosine similarity, was used [51]. Values of the Tanimoto scores near 0 indicate reactions with no or negligible similarity, whereas values near 1 indicate reactions with high similarity.

The Tanimoto score for each descriptive layer, T_{Lk}, together with the global Tanimoto score, T_G, was calculated. The Tanimoto score for the k-th descriptive layer was defined as:

Equation 4.1: The Tanimoto score for the k-th layer $$T_{Lk} = \frac{c_k}{a_k + b_k - c_k}$$

where a_k was the count of the fragments in the k-th layer of reaction fingerprint A; b_k was the count of the fragments in the k-th layer of reaction fingerprint B; and c_k was the number of common k-th layer fragments of reaction fingerprints A and B. Two fragments were equal if their canonical SMILES and their stoichiometric coefficients were identical. The global Tanimoto similarity score, T_G, was defined as follows:

Equation 4.2: The global Tanimoto similarity score $$T_G = \frac{\sum_{k=0}^{7} c_k}{\sum_{k=0}^{7} a_k + \sum_{k=0}^{7} b_k - \sum_{k=0}^{7} c_k}$$

For each reaction fingerprint, its Tanimoto similarity score was calculated against the reaction fingerprints from the BridgIT reference database, which contained reaction fingerprints of all known, well-characterized enzymatic reactions (Figure 4.1, panel 3).

4.2.1.4 Sorting, ranking and gene assignment

For a given input reaction, the reference reactions (Explained in the section 4.2.2) were ranked using the computed T_G scores. The algorithm distinguished between the identified reference reactions with the same T_G score based on the T_L score of layers 0 and 1, and it also allows the user to assign ranking weights to specified layers. The protein sequences associated with the highest ranked, i.e., the most similar, reference reactions were then assigned to the input reaction (Figure 4.1, panel 4).

4.2.2 Reference reaction database

The BridgIT reference reaction database is an essential component of the BridgIT workflow (Figure 4.1). It consists of well-characterized reactions with associated genes and protein sequences, and it was built based on the KEGG 2016 reaction database. The KEGG database is the most comprehensive database of enzymatic reactions, and it provides information about biochemical reactions together with their corresponding enzymes and genes. However, half of KEGG reactions lack associated genes and protein sequences, and they are hence considered to be orphan reactions. The BridgIT reference database was built using the KEGG reactions that (i) can be reconstructed by the existing BNICE.ch generalized reaction rules and are elementally balanced (5,270 reactions) and (ii) are non-orphan (5,049 reactions). This restriction removes reactions that lack characterized substrate reactive sites, meaning that they cannot be used in our comparisons. As a result, the reference reaction database contains information for 5,049 out of 9,556 KEGG reactions.

4.3 Results and Discussion

4.3.1 Sensitivity analysis of the BridgIT fingerprint size

The defining characteristic of the BridgIT reaction fingerprint is that it is centered around the reactive site of the reaction substrate(s). The number of description layers in the BridgIT fingerprint, i.e., the fingerprint size, defines how large of a chemical structure around the reactive site we consider when evaluating the similarity (See section 4.2.1.2). To investigate to what extent the fingerprint size affects the similarity results, we performed a sensitivity analysis where we varied the fingerprint size between 0 to 10.

For this analysis, we considered the 5,049 non-orphan KEGG reactions that existed in the BridgIT reference reaction database. We started by forming reaction fingerprints that contained only the description layer 0 (fingerprint size 0) and evaluated how many of 5,049 non-orphan reactions BridgIT could correctly identify.

We next formed the reaction fingerprints using only the description layers 0 and 1 (fingerprint size 1), and we performed the evaluation again. We repeated this procedure until the final step, where we formed the reaction fingerprints with ten description layers (fingerprint size 10).

As expected, the increase in the fingerprint size, i.e., specificity, led to a decrease in the average number of similar reactions assigned to the studied reactions. Moreover, the more description layers that were incorporated into the BridgIT fingerprint, the more accurately BridgIT matched the analyzed reactions (Table 4.1). Already for a fingerprint size 7, BridgIT correctly mapped 100% of the analyzed reactions, i.e., each of the 5,049 non-orphan reactions was matched to itself in the reference reaction database. This indicated that the information about chains of eight atoms along with their connecting bonds around the reactive sites was sufficient for BridgIT to correctly match all non-orphan KEGG reactions, and we chose the fingerprint size 7 for our further studies.

Table 4.1: Percent of correctly mapped reactions as a function of the size of the BridgIT and the standard fingerprint.

	Fingerprint size										
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
BridgIT fingerprint % correctly mapped reactions	4.3	35.2	60.5	72.1	92.7	97.8	98.6	**100**	100	100	100

4.3.2 BridgIT reaction fingerprints offer improved predictions

To evaluate BridgIT performances against existing approaches in this field [40], [42, p.], [52], we performed two comparative studies. In the first study, we repeated the analysis from the previous section using the standard reaction difference fingerprint (Section 4.2.1.2), which is used in structure similarity methods such as RxnSim [38] and RxnFinder [39], to assess the benefits of introducing the information about the reactive site of substrates into the reaction fingerprints. A comparison of the two sets of predictions on 5,049 non-orphan reactions showed that the predictions obtained with BridgIT-modified fingerprints were significantly better than the standard ones. BridgIT identified 100% of non-orphan reactions correctly versus the 71% success rate for the standard fingerprint method. Furthermore, BridgIT correctly matched 93% of the analyzed enzymatic reactions using the information about only four connecting bonds around the atoms of the reactive sites (fingerprint size 4) (Table 4.1), which exceeds the 71% of matched reactions when using the standard reaction fingerprints (fingerprint size 7).

The inferior performance of the standard reaction fingerprint method arose from three main sources. First, fragments from the substrate and product sets were cancelled out upon algebraic summation inside the

fingerprint description layers (Section 4.2.1.2), in which description layers 0 and 1 define the single atoms and the connected pairs of atoms of the reactive site, and layers 2 to 7 include information about the chemical structure around the reactive site that contains up to eight atoms and seven bonds (Figure 4.1). This cancellation occurred in all description layers (fingerprint size 7) for 246 non-orphan reactions, i.e., their standard fingerprints were empty. As an example, Figure 4.2 shows the standard reaction fingerprint of KEGG reaction R00722 that was empty for the standard fingerprint method. The information about reactive sites introduced in the BridgIT reaction fingerprints prevents such cancellations, since BridgIT does not include the atoms of the reactive site(s) in the process of the algebraic summation of the substrate and product set fragments (Section 4.2.1.2). As a result, BridgIT mapped R00722 to itself and identified R00330 as the most similar reaction to R00722 (Figure 4.2, panel A). Indeed, according to the KEGG database, the enzyme 2.7.4.6 catalyzes both reactions.

Second, the performance of the standard reaction fingerprint suffered because the first description layer of the standard fingerprint was empty for an additional 1,129 reactions, which indicated that these fingerprints did not represent the bond changes during the reaction.

Third, the remaining 89 mismatched non-orphan reactions had partial cancellations in the fingerprint description layers. For example, the standard fingerprint method incorrectly identified R03132 as the most similar to R00691, whereas BridgIT identified R00691 and R01373 as the most similar to R00691 (Figure 4.2, panel B), which matches the KEGG reports indicating that both R00691 and R01373 can be catalyzed by either EC 4.2.1.51 or EC 4.2.1.91.

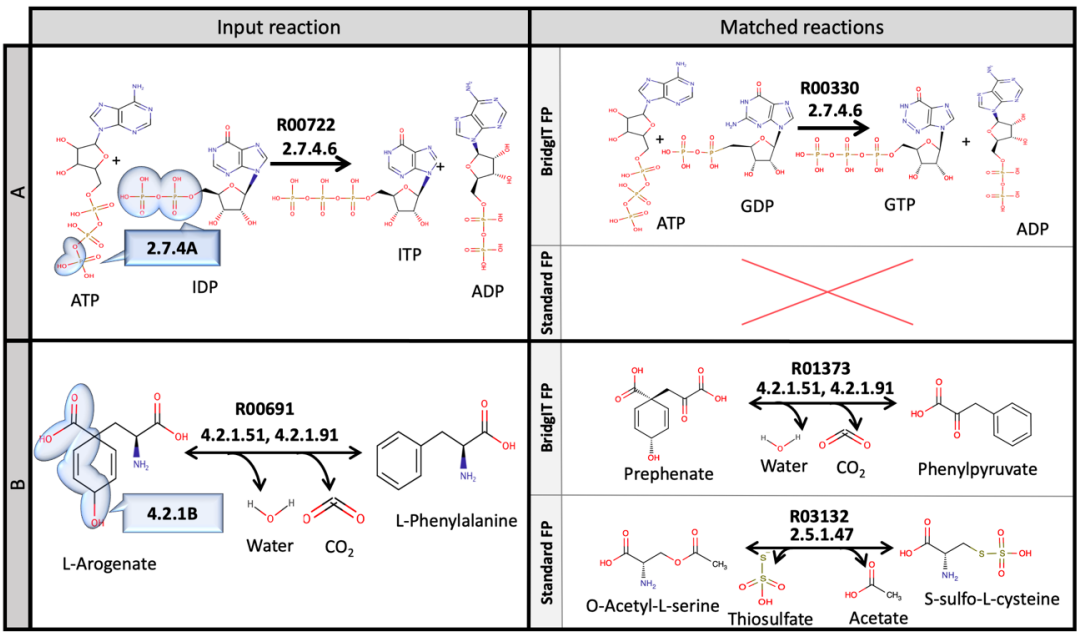


Figure 4.2: Comparison of the results obtained with the BridgIT and standard fingerprint on two example KEGG reactions. (A) The input reaction R00722 (left) and the most similar reactions (right) identified with the BridgIT and standard fingerprints. Note that the

standard fingerprinting method failed to find a similar reaction to R00722 due to cancellations inside all fingerprint description layers. (B) The input reaction R00691 (left) and the most similar reactions (right) identified with the BridgIT and standard fingerprints.

In the second study, we compared the performance of BridgIT method against three state-of-art methods EC-BLAST (42), Selenzyme (48), and E-zyme2 (40) on two benchmark problems. The benchmark problems consisted of identifying the most similar reactions to two example reactions each representing a class of reactions that appear ubiquitously in biochemical networks. We choose R00722 (Figure 4.2, panel A) to exemplify the first class of reactions characterized by a very similar structure of substrates and products, and R07500 to represent the class of multi-substrate multi-product reactions (Appendix, Table 8.2 and Table 8.3). For the two benchmark reactions, we ranked the similar reactions proposed by each of methods according to the corresponding similarity scores, and top 100 similar reactions proposed by each method were used for comparisons.

The most similar reaction proposed by BridgIT correctly matched the 4th level EC number (2.7.4.6) of the first benchmark reaction R00722 (Appendix, Table 8.2). Three out of four EC-BLAST variants [42] proposed a set of the reactions with the maximal similarity score (Appendix, Table 8.2). This set contained reactions that correctly matched the 4th level EC number of R00722, but also reactions with EC numbers not even matching the 1st level EC number of the benchmark reaction (Appendix, Table 8.2). The three variants of Selenzyme [52] proposed reactions that could match only the 3rd level EC number of R00722, whereas E-zyme2 was unable to find a matching reaction due to very similar structures in the substrate-product pairs (Appendix, Table 8.2).

In the second benchmark, none of the investigated methods could propose reactions that match the EC number of R07500 (2.5.1.115) up to the 4th level, and all methods could match the 3rd level EC number for this reaction (Appendix, Table 8.3). BridgIT proposed 39 similar reactions matching the 3rd level EC numbers of R07500, whereas the EC-BLAST variant with structural similarity proposed 45, Selenzyme 10, E-zyme2 9, and the three other EC-BLAST variants proposed 5-7 such reactions (Appendix, Table 8.3). Additionally, we performed receiver operating characteristic (ROC) analysis on the sets of proposed similar reactions, and out of all compared methods BridgIT had the highest area under the curve (AUC) index of 0.95, meaning that it had the best performance among the compared methods for this class of reactions (Appendix, Table 8.3).

The results of these two studies demonstrate the potential of BridgIT to outperform the currently available methods for enzyme annotation.

4.3.3 From reaction chemistry to detailed enzyme mechanisms

Approximately 15% of KEGG reactions (1,532 reactions) are assigned to more than one enzyme and EC number, i.e., multiple enzymes can catalyze a specific biotransformation through different enzymatic mechanisms. For example, KEGG reaction R00217 is assigned to three different EC numbers, 4.1.1.3

(oxaloacetate carboxy-lyase), 1.1.1.40, and 1.1.1.38 (both malate dehydrogenases), and the corresponding reactions involve different mechanisms (Figure 4.3). The reaction mechanism of the 4.1.1.3 enzyme is well understood, as it belongs to the carboxy-lyases, where a carbon-carbon bond is broken and a molecule of CO_2 is released. This enzyme can decarboxylate three different compounds: glutaconyl-CoA, methylmalonyl-CoA, and oxaloacetate (from this example). The overlapping reactive site of these three compounds is captured in the 4.1.1B rule of BNICE.ch (Figure 4.3, panel C). In contrast, the 1.1.1.38 enzyme found in bacteria and insects and 1.1.1.40 found in fungi, animals, and plants are rather specific enzymes that decarboxylate oxaloacetate and malate with two different mechanisms. The decarboxylation is performed in the case of oxaloacetate without and in the case of malate with the incorporation of NAD+ as a cofactor. The difference in the structure of these two molecules is only in having either a ketone or an alcohol group on the second carbon. Consequently, the structure of the reactive site that these enzymes recognize has to reflect the difference between malate and oxaloacetate, and this is well captured in the 1.1.1A rule of BNICE.ch. The 4.1.1B rule requires a less specific reactive site compared to the 1.1.1A rule, and these two rules have two different reaction fingerprints for catalyzing the same reaction R00217 because they describe different mechanisms for the same reaction.

Moreover, for 42% of the KEGG reactions that have a single enzyme assigned to them, BNICE.ch identified multiple alternative reactive sites and created multiple reaction fingerprints that describe the biotransformation of these reactions. Therefore, a single reaction from KEGG was translated into more than one fingerprint in the BridgIT reference database. This way, by preserving the information about enzyme binding pockets, the reconstructed BridgIT reference reaction database expands from 5,049 *reactions* to 17,657 *reaction fingerprints* corresponding to 17,657 *detailed reaction mechanisms*.

Currently, BridgIT is the only method that can distinguish different reaction mechanisms for the reactions catalyzed by different enzymes. As a consequence, BridgIT can propose distinct sets of protein sequences corresponding to distinct mechanisms and rank them according to the BridgIT score. The protein sequences can then be prioritized based on the BridgIT ranking, enzyme specificity, and the host organism.

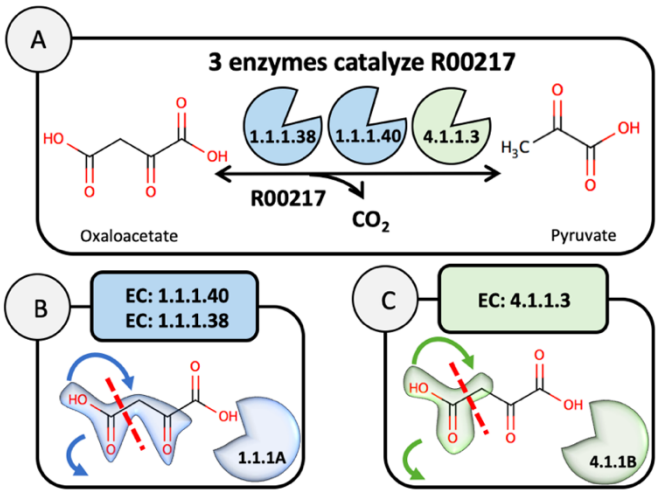


Figure 4.3: A multi-enzyme reaction such as R00217 can be catalyzed by more than one enzyme. BridgIT identified two distinct fingerprints for this reaction that correspond to two reactive sites of oxaloacetate. The reactive site recognized by the 1.1.1.- rule is more specific (blue substructure) than the one recognized by the 4.1.1.- rule (green substructure).

4.3.4 Comparison of BridgIT and BLAST predictions

As a means to relate *reaction structural similarity* obtained using BridgIT with *reaction sequence similarity* obtained using BLAST [53], we applied these two techniques in parallel on a subset of reactions and their corresponding protein sequences from the reference reaction database. We compared the similarity results of BridgIT with those of BLAST, and we statistically assessed BridgIT performance using receiver operating characteristic (ROC) curve analysis.

We chose *E. coli* BW29521 (EBW) as our benchmark organism for this analysis. There were 531 non-orphan reactions in EBW associated with 413 protein sequences. In total, there were 731 reaction-gene associations, as there were reactions with more than one associated gene, and genes associated with more than one reaction. We removed all of the non-orphan reactions of EBW from the BridgIT reference database and we removed their associated protein sequences from the KEGG protein sequence database. We then used BridgIT to assess the structural similarity of the 531 EBW reactions to the BridgIT reference reactions using the Tanimoto score, and we also applied BLAST to quantify the similarity of the 413 EBW protein sequences to the KEGG protein sequence database using e-values. The concept of the validation procedure is illustrated in Figure 4.4. We provided a list of BridgIT reaction-reaction comparisons together with BLAST sequence-sequence comparisons.

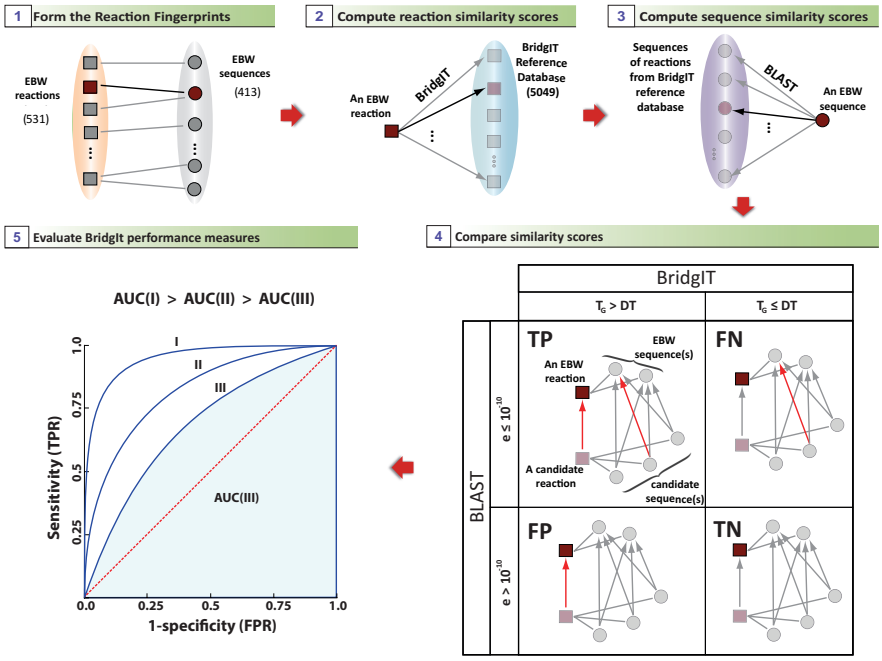


Figure 4.4: Five steps in the BridgIT cross validation procedure.

Comparing reaction (BridgIT) and sequence (BLAST) similarity scores. We considered two sequences to be similar if BLAST reported an e-value of less than 10^{-10} for their alignment. For a chosen discrimination threshold (DT) of the global Tanimoto score (T_G) we considered the BridgIT prediction of similarity between an EBW reaction and a BridgIT reference reaction with a Tanimoto score of T_G as:

I. True Positive (TP) if $T_G > DT$ and their associated sequence(s) were similar (e-value < 10^{-10});
II. True Negative (TN) if not similar for both BridgIT ($T_G < DT$) and BLAST+ (e-value > 10^{-10});
III. False Positive (FP) if similar for BridgIT ($T_G > DT$) but not similar for BLAST+ (e-value > 10^{-10});
IV. False Negative (FN) if not similar for BridgIT ($T_G < DT$) but similar for BLAST+ (e-value < 10^{-10}).

We then counted the number of TPs, TNs, FPs, and FNs for all 531 reactions, and we summed these quantities to obtain the total number of TPs, TNs, FPs, and FNs per chosen DT. We repeated this procedure for a set of DT values varying across the interval between 0 and 1. Finally, we used the total number of TPs, TNs, FPs, and FNs to compute the true positive and false positive rates for the ROC curve analysis (Figure 4.5, panel A). The ROC curve indicated that the reaction comparison based on *reaction structural similarity* (BridgIT) was comparable to the one based on *reaction sequence similarity* (BLAST). Indeed, the obtained area under the ROC curve (AUC) score for the BridgIT classifier was 0.91, indicating that the similarities between the two methods were very high (Figure 4.5, panel A). We next studied if the type of compared reactions affected the accuracy of BridgIT predictions by categorizing reactions according to their first-level EC class, which indicates the broadest category of enzyme functionality, and then performing the ROC analysis for each class separately (Figure 4.5, panel A). The analysis revealed that BridgIT performed well with all major enzyme classes, as represented by the high AUC scores, ranging from 0.88 (EC 1) to 0.96 (EC 5).

We next analyzed the accuracy of BridgIT classification as a function of the DT of the Tanimoto score (Figure 4.5, panel B). The accuracy ranged from 43% for DT = 0.01 to 85% for DT = 0.30. For values of DT > 0.30, the accuracy monotonically decreased toward a value of 62% for DT = 1. The classifier was overly conservative for values of DT > 0.30, and it was rejecting true positives (Figure 4.5, panel B). More specifically, for DT = 0.30, the TP percentage was 38%, whereas, for DT = 1, it was reduced to 3%. In contrast, the TN percentage increased very slightly for the values of DT > 0.30, where for DT = 0.30, it was 46%, and for DT = 1, it was 57% (Figure 4.5, panel A). Based on this analysis, we have chosen a DT of 0.30 as an optimal threshold value for further studies.

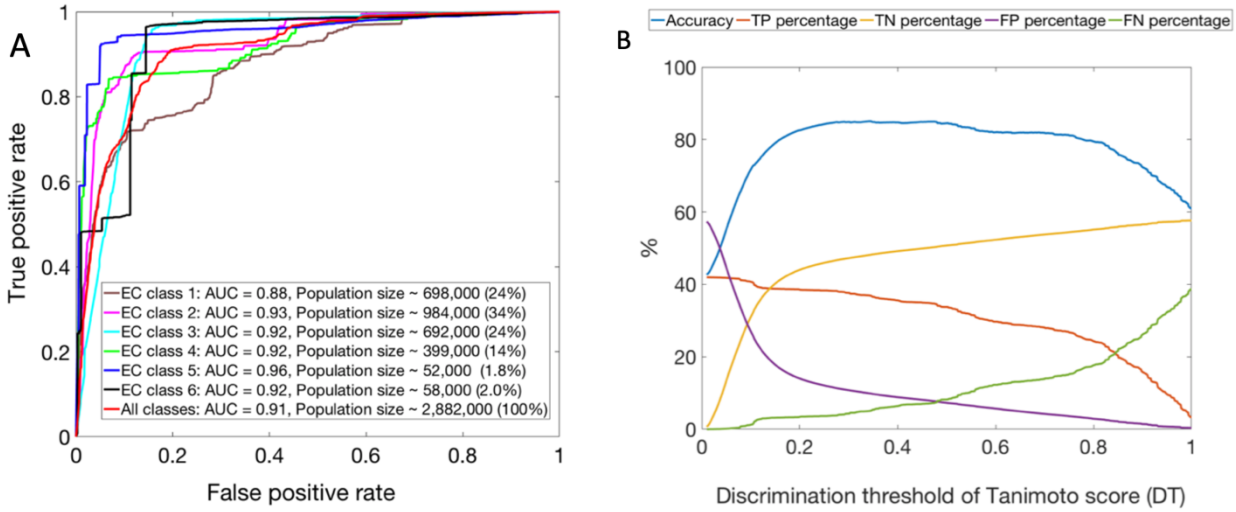


Figure 4.5: Panel A: ROC curve for the BridgIT classifier among all EC classes and inside each class. Panel B: Accuracy characteristics and the percentages of TP, TN, FP and FN as a function of the discrimination threshold DT. The percentages are computed as X %=100*X/(TP+TN+FN+FP) where X can be TP, TN, FP or FN.

4.3.5 BridgIT analysis of known reactions with common enzymes

The 5,049 reactions in the reference database were catalyzed by only 2,983 enzymes, i.e., there were promiscuous enzymes that catalyzed more than one reaction. Out of the 2,983 enzymes, 844 of them were promiscuous, catalyzing 2,432 of the reactions. Interestingly, BridgIT correctly assigned more than 80% of these 2,432 reactions to their corresponding promiscuous enzyme. An example of such a group is given in

Table 4.2. This table shows the same enzymes listed across the top and down the size of the grid, with the corresponding Tanimoto scores indicating the accuracy of BridgIT's classifications. The overall high scores in this grid indicate the accuracy of the enzyme assignments.

We investigated the remaining 20% of reactions in depth, and we observed that the Tanimoto scores of the first two description layers (Section 4.3.1.2) indicated a very low similarity between the reactions catalyzed by the same enzyme. This result suggested that such enzymes were either multi-functional, i.e., they had more than one reactive site (Figure 4.6), or were incorrectly classified in the EC classification system.

Table 4.2: A group of five reactions catalyzed by enzyme 1.1.1.219, wherein the Tanimoto score is given for the comparison between the reaction listed across the top and the reaction listed down the side.

Catalyzed reactions	R03123	R03636	R05038	R07999	R07998
R03123	1	0.96	0.93	0.93	0.98
R03636	0.96	1	0.96	0.94	0.95

R05038	0.93	0.96	1	0.97	0.91
R07999	0.93	0.94	0.97	1	0.91
R07998	0.98	0.95	0.91	0.91	1

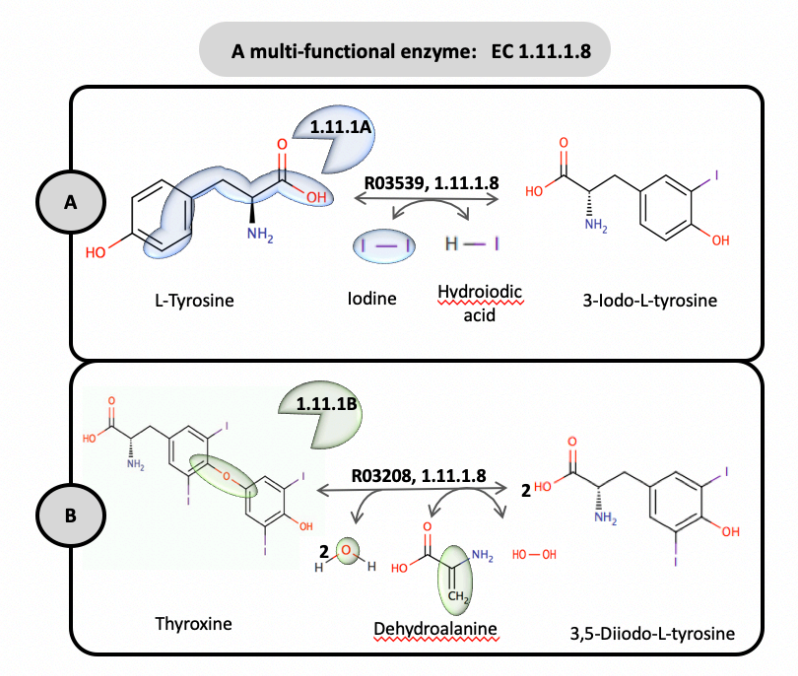


Figure 4.6: Multi-functional enzymes can catalyze reactions with two different reactive sites. (A) R03539 and (B) R03208 are catalyzed by the same enzyme, 1.11.1.8. However, the reactive sites of these substrates are completely different

4.3.6 BridgIT validation against biochemical assays

To assess BridgIT's performance using biochemically confirmed reactions, we performed two validation studies on sets of (I) orphan and (II) novel reactions. Since the known reactions in KEGG are all experimentally confirmed using biochemical assays, we could use this pooled experimental data from hundreds of laboratories to demonstrate BridgIT's ability to identify potential enzymes for catalyzing the biologically relevant orphan reactions on a large scale.

Study I: We compared the number of orphan reactions in the two versions of the KEGG reaction database, KEGG 2011 and KEGG 2018. We found that 234 orphan reactions from KEGG 2011 were later associated with enzymes in KEGG 2018, meaning they became non-orphan reactions. Since these newly classified reactions have been experimentally confirmed, we used these 234 reactions as a benchmark to evaluate BridgIT performance.

We formed the reference reaction database using the reactions from KEGG 2011 (Section 4.2.2), and we compared the BridgIT results with the KEGG 2018 enzyme assignments up to the third EC level. Remarkably, BridgIT and KEGG 2018 assigned enzymes matched to the third EC level for 211 out of 234 (90%) reactions. This means that BridgIT accurately predicted the enzyme mechanism for enzymes that have been biochemically confirmed to catalyze a large majority of the orphan reactions in 2011. In addition, the set of protein sequences proposed by BridgIT comprised highly related protein sequences to the ones assigned to these enzymes in KEGG 2018.

The 234 reactions are catalyzed by 168 enzymes with specified fourth-level EC numbers in KEGG 2018. However, only 29 out of these 168 enzymes were cataloged in KEGG 2011, and the remaining 139 enzymes had new fourth-level EC classes assigned in KEGG 2018 – meaning BridgIT only had access to the 29 enzymes that were classified in KEGG 2011 from which the reference reaction database was built. The 29 enzymes catalyzed 35 out of the 234 studied reactions. For 29 out of these 35 (83%) orphan reactions, the BridgIT algorithm predicted the same sequences that KEGG 2018 assigned to these reactions. A higher matching score when comparing up to the third EC level rather than the fourth EC level is likely because BridgIT uses BNICE.ch generalized reaction rules, which describe the biotransformations of reactions with specificities up to the third EC level.

***Study II*:** The ATLAS of biochemistry [46] provides a comprehensive catalog of theoretically possible bio-transformations between KEGG compounds, and it can be mined for novel biosynthetic routes for a wide range of applications in metabolic engineering, synthetic biology, drug target identification, and bioremediation (40)**.** We studied the 379 reactions from the ATLAS of Biochemistry that were novel in KEGG 2014 and were later experimentally identified and catalogued in KEGG 2018.
We formed the reference reaction database using the reactions from KEGG 2014 and applied BridgIT to these 379 reactions. For 334 out of these 379 reactions, BridgIT proposed similar known reactions with a Tanimoto score higher than 0.3, thus providing promising protein sequences for enzymes catalyzing these reactions. For 14 of these novel reactions, BridgIT assigned the same sequences that were assigned in KEGG 2018. An example of such a reaction is rat132341, which was a novel reaction in 2014 and later was catalogued as R10392 in KEGG 2018 (Figure 4.7, panel A). The BridgIT analysis of this reaction revealed that R03444, which is catalyzed by enzyme 4.2.1.114, is the structurally closest reaction to this novel one, suggesting that protein sequences from EC 4.2.1.114 can catalyze this novel reaction. This was later confirmed by experimental biochemical evidence, as R10392 is associated with the same EC 4.2.1.114 enzyme in KEGG 2018. There are 243 available protein sequences for enzyme 4.2.1.114, and one sequence already has a confirmed protein structure (Figure 4.7, panel C). This represents the first computational method for predicting protein sequences for orphan and novel reactions whose results were validated using experimental biochemical evidence on a large scale.

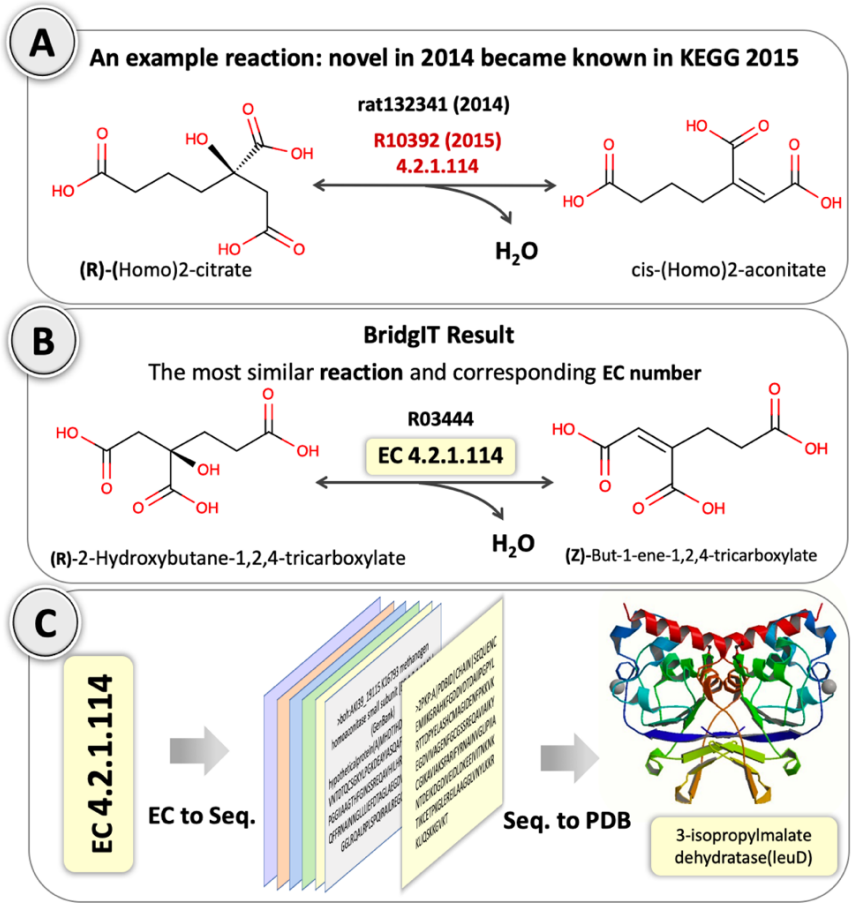


Figure 4.7: Details of the BridgIT verification procedure that was performed on ATLAS reaction rat132341, which was novel in KEGG 2014 and later experimentally identified and catalogued in KEGG 2018 — i.e., it became a non-orphan reaction (R10392). (A) rat132341 catalyzes the conversion of (R)-(Homo)2-citrate to cis-(Homo)2-aconitate. (B) Using the biochemical knowledge of KEGG 2014, BridgIT predicts the KEGG reaction R03444, which is catalyzed by a 4.2.1.114-class enzyme, as the most similar known reaction to rat132341. Remarkably, the same enzyme is later assigned to R10392 in KEGG 2018 with the corresponding biochemical confirmation. (C) The identified EC number (4.2.1.114) can be used to extract the corresponding protein sequences along with their crystal structures.

4.3.7 BridgIT predictions for KEGG 2018 orphan reactions

We applied BridgIT to the 810 orphan KEGG 2018 reactions that could be reconstructed using the BNICE.ch generalized reaction rules. The remaining 1646 orphan reactions could not be reconstructed because they are either not balanced or lack the structure for at least one of their substrates. Remarkably, BridgIT identified corresponding reference reactions with Tanimoto scores higher than the optimal threshold value of 0.30 for 97% of the orphan reactions. The remaining 3% of orphan reactions had a low similarity with the reference reactions. A large number of the orphan reactions originate from the pathways toward plant and microbial natural products that frequently involve complex and less-investigated classes of enzymes such as polyketide

synthases (PKS), non-ribosomal peptide synthetases (NRPSs), Terpene Cyclases (TCs) and Cytochromes P450s. Interestingly, BridgIT mapped 112 out of 810 orphan reactions back to these families, i.e., it predicted that 72 orphan reactions can be catalyzed by P450s, 33 by PKSs, 6 by NRPSs and 1 by TC.

This result and the fact that BridgIT correctly mapped 100% of non-orphan KEGG reactions suggested that, as our knowledge of biochemistry expands, the annotation of novel and orphan reactions using tools such as BridgIT will also improve.

4.3.8 BridgIT predictions for ATLAS novel reactions

We further utilized BridgIT to identify candidate enzymes for all the 137,000 *de novo*, orphan-like, ATLAS reactions. These candidate enzymes can either be used directly in systems biology designs if the matched enzymes perform the desired catalysis, or their amino acid sequences can be optimized through protein engineering to achieve the desired results. We found that 7% of novel ATLAS reactions were matched to known KEGG reactions with a Tanimoto score of 1 (perfect match), while 88% were similar to KEGG reactions with a Tanimoto score higher than the optimal threshold value of 0.3. Therefore, BridgIT could identify promising enzyme sequences for catalyzing 95% of novel ATLAS reactions. The remaining 5% of these reactions were not similar to any of the well-characterized, known enzymatic reactions.

Finding well-characterized reactions that are similar to novel ones is crucial for evolutionary protein engineering as well as computational protein design, and methods like BridgIT can be instrumental in moving from a concept to the experimental implementation of *de novo* reactions. Additionally, to facilitate the experimental implementation of novel ATLAS reactions in metabolic engineering, systems and synthetic biology, and bioremediation studies, we can use the BridgIT similarity scores as confidence measures for evaluating the feasibility.

The results of the BridgIT analysis of the KEGG 2018 orphan and novel ATLAS reactions are available on the website http://lcsb-databases.epfl.ch/atlas/.

4.4 Access to online version of BridgIT

The online version of BridgIT is available on the homepage of LCSB website (http://lcsb-databases.epfl.ch). Users can access BridgIT after registration and creating account.

In the analyse tab of BridgIT page, users can upload a zip file including: a text file of reaction equations (one reaction per line) and one folder containing the molfiles of reaction participants. For more details about the format of the input file, please check the user manual of BridgIT available on LCSB website.

After the successful uploading of input the latest version of BridgIT will be running in the background and users will receive a link corresponding to the submitted input. Depending on the complexity and also number

of reactions in the input file results will be ready in some minutes to some hours. The results can be directly accessed from download link and are available for one week, after that they will be automatically deleted.

In the result file, the most similar metabolic reactions to input reaction are ranked based on their BridgIT score. Also, each similar metabolic reaction is annotated with it's EC number and this EC number is used to get gene information from protein sequence databases.

4.5 Conclusion and Outlook

We developed the computational tool, BridgIT, to evaluate and quantify the structural similarity of biochemical reactions by exploiting the biochemical knowledge of BNICE.ch generalized reaction rules. Because the generalized reaction rules can identify reactive sites of substrates, BridgIT can translate the structural definition of biochemical reactions into a novel type of reaction fingerprint that explicitly describes the atoms of the substrates' reactive sites and their surrounding structure. Through the analysis of 5,049 known and well-defined biochemical reactions, we found that knowledge of the neighborhood up to three bonds away from the atoms of the reactive site can predict biochemistry and match catalytic protein sequences. The reaction fingerprints proposed in this work can be used to compare all novel and orphan reactions to well-characterized reference reactions and, consequently, to link them with genes, genomes, and organisms. We demonstrated through several examples the improvements that the BridgIT fingerprint brings to the field compared to the fingerprints currently existing in the literature.

A drawback of traditional sequence similarity methods is that they cannot identify protein sequence candidates for *de novo* reactions, which we have shown BridgIT can do.

We tested BridgIT predictions against experimental biochemical evidence, within two large-scale validations studies on sets of (i) 234 orphan and (ii) 379 *de novo* reactions. The reactions from these two sets were unknown in the previous versions of the KEGG database but were later experimentally confirmed and catalogued in KEGG 2018. BridgIT predicted the exact or a highly related enzyme for 89% of these reactions.

We further applied BridgIT to the entire catalog of *de novo* reactions of the ATLAS of Biochemistry database and proposed several candidate enzymes for each of them. The candidate enzymes for these *de novo* reactions are either immediately capable of catalyzing these reactions or can serve as initial sequences for enzyme engineering. The obtained BridgIT similarity scores can also be used as a confidence score to assess the feasibility of the implementation of novel ATLAS reactions in metabolic engineering and systems biology studies.

The applications of BridgIT go beyond merely bridging gaps in metabolic reconstructions, as this method can be used to identify the potential utility of existing enzymes for bioremediation as well as for various applications in synthetic biology and metabolic engineering. As the field of metabolic engineering grows and

metabolic engineering applications increasingly turn towards the production of valuable industrial chemicals such as 1,4-butanediol [54], [55], we expect that methods for the design of *de novo* synthetic pathways, such as BNICE.ch [16], and methods for identifying candidate enzymes for *de novo* reactions, such as BridgIT, will grow in importance.

4.6 References

[1] J. D. Orth *et al.*, "A comprehensive genome-scale reconstruction of Escherichia coli metabolism—2011," *Mol. Syst. Biol.*, vol. 7, no. 1, p. 535, Jan. 2011, doi: 10.1038/msb.2011.65.

[2] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, "KEGG: new perspectives on genomes, pathways, diseases and drugs," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D353–D361, Jan. 2017, doi: 10.1093/nar/gkw1092.

[3] M. Sorokina, M. Stam, C. Medigue, O. Lespinet, and D. Vallenet, "Profiling the orphan enzymes," *Biol. Direct*, p. 9, 2014.

[4] A. G. Shearer, T. Altman, and C. D. Rhee, "Finding Sequences for over 270 Orphan Enzymes," *PLoS ONE*, vol. 9, no. 5, p. e97250, May 2014, doi: 10.1371/journal.pone.0097250.

[5] J. Gao, L. B. M. Ellis, and L. P. Wackett, "The University of Minnesota Biocatalysis/Biodegradation Database: improving public access," *Nucleic Acids Res.*, vol. 38, no. suppl_1, pp. D488–D491, Jan. 2010, doi: 10.1093/nar/gkp771.

[6] V. Hatzimanikatis, C. H. Li, J. A. Ionita, and C. S. Henry, "Exploring the diversity of complex metabolic networks," *Bioinformatics*, vol. 21, pp. 1603–1609, 2005.

[7] V. Hatzimanikatis, C. Li, J. A. Ionita, and L. J. Broadbelt, "Metabolic networks: enzyme function and metabolite structure," *Curr. Opin. Struct. Biol.*, vol. 14, no. 3, pp. 300–306, Jun. 2004, doi: 10.1016/j.sbi.2004.04.004.

[8] K. C. Soh and V. Hatzimanikatis, "DREAMS of metabolism," *Trends Biotechnol.*, vol. 28, no. 10, pp. 501–508, Oct. 2010, doi: 10.1016/j.tibtech.2010.07.002.

[9] P. Carbonell, A.-G. Planson, D. Fichera, and J.-L. Faulon, "A retrosynthetic biology approach to metabolic pathway design for therapeutic production," *BMC Syst. Biol.*, vol. 5, no. 1, p. 122, 2011, doi: 10.1186/1752-0509-5-122.

[10] G. Rodrigo, J. Carrera, K. J. Prather, and A. Jaramillo, "DESHARKY: automatic design of metabolic pathways for optimal cell growth," *Bioinformatics*, vol. 24, no. 21, pp. 2554–2556, Nov. 2008, doi: 10.1093/bioinformatics/btn471.

[11] A. Cho, H. Yun, J. Park, S. Lee, and S. Park, “Prediction of novel synthetic pathways for the production of desired chemicals,” *BMC Syst. Biol.*, vol. 4, no. 1, p. 35, 2010, doi: 10.1186/1752-0509-4-35.

[12] H. Yim, R. Haselbeck, W. Niu, and C. Pujol-Baxley, “Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol,” *Nat Chem Biol*, pp. 445–452, 2011.

[13] M. A. Campodonico, B. A. Andrews, J. A. Asenjo, B. O. Palsson, and A. M. Feist, “Generation of an atlas for commodity chemical production in Escherichia coli and a novel pathway prediction algorithm, GEM-Path,” *Metab. Eng.*, vol. 25, pp. 140–158, 2014.

[14] K. L. J. Prather and C. H. Martin, “De novo biosynthetic pathways: rational design of microbial chemical factories,” *Curr. Opin. Biotechnol.*, vol. 19, no. 5, pp. 468–474, Oct. 2008, doi: 10.1016/j.copbio.2008.07.009.

[15] B. Delépine, T. Duigou, P. Carbonell, and J.-L. Faulon, “RetroPath2.0: A retrosynthesis workflow for metabolic engineers,” Jun. 2017, doi: 10.1101/141721.

[16] N. Hadadi and V. Hatzimanikatis, “Design of computational retrobiosynthesis tools for the design of de novo synthetic pathways,” *Curr Opin Chem Biol*, vol. 28, pp. 99–104, 2015.

[17] P. Carbonell, P. Parutto, J. Herisson, S. B. Pandit, and J. L. Faulon, “XTMS: pathway design in an eXTended metabolic space,” *Nucleic Acids Res.*, vol. 42, pp. 389–394, 2014.

[18] N. Hadadi, K. C. Soh, M. Seijo, and A. Zisaki, “A computational framework for integration of lipidomics data into metabolic pathways,” *Metab. Eng.*, vol. 23, pp. 1–8, 2014.

[19] O. Rolfsson, B. Ø. Palsson, and I. Thiele, “The human metabolic reconstruction Recon 1 directs hypotheses of novel human metabolic functions,” *BMC Syst. Biol.*, vol. 5, no. 1, p. 155, 2011, doi: 10.1186/1752-0509-5-155.

[20] P. D. Karp, “Call for an enzyme genomics initiative,” *Genome Biol.*, vol. 5, 2004.

[21] J. D. Orth and B. O. Palsson, “Systematizing the Generation of Missing Metabolic Knowledge,” *Biotechnol. Bioeng.*, vol. 107, pp. 403–412, 2010.

[22] A. Osterman and R. Overbeek, “Missing genes in metabolic pathways: a comparative genomics approach,” *Curr Opin Chem Biol*, vol. 7, pp. 238–251, 2003.

[23] R. Overbeek, M. Fonstein, M. D’Souza, G. D. Pusch, and N. Maltsev, “The use of gene clusters to infer functional coupling,” in *Proceedings of the National Academy of Sciences of the United States of America 1999, 96*, pp. 2896–2901.

[24] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates, “Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles,” in *Proceedings of the National Academy of Sciences of the United States of America 1999, 96*, pp. 4285–4288.

[25] V. Chen, “Predicting genes for orphan metabolic activities using phylogenetic profiles,” *Genome Biol*, p. 17, 2006.

[26] R. Overbeek, T. Begley, R. M. Butler, and J. V. Choudhuri, “The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes,” *Nucleic Acids Res*, vol. 33, pp. 5691–5702, 2005.

[27] D. Vallenet, L. Labarre, Z. Rouy, and V. Barbe, “a microbial genome annotation system supported by synteny results,” *Nucleic Acids Res.*, vol. 34, pp. 53–65, 2006.

[28] P. Kharchenko, L. F. Chen, Y. Freund, D. Vitkup, and G. M. Church, “Identifying metabolic enzymes with multiple types of association evidence,” *BMC Bioinformatics*, p. 7, 2006.

[29] Y. Yamanishi, H. Mihara, M. Osaki, and H. Muramatsu, “Prediction of missing enzyme genes in a bacterial metabolic network - Reconstruction of the lysine-degradation pathway of Pseudomonas aeruginosa,” *Febs J.*, vol. 274, pp. 2262–2273, 2007.

[30] Y. Chen, F. L. Mao, G. Li, and Y. Xu, “Genome-wide discovery of missing genes in biological pathways of prokaryotes,” *BMC Bioinformatics*, p. 12, 2011.

[31] A. A. T. Smith, E. Belda, A. Viari, C. Medigue, and D. Vallenet, “The CanOE Strategy: Integrating Genomic and Metabolic Contexts across Multiple Prokaryote Genomes to Find Candidate Genes for Orphan Enzymes,” *Plos Comput Biol*, vol. 8, 2012.

[32] W. R. Pearson, “An Introduction to Sequence Similarity (‘Homology’) Searching,” *Curr. Protoc. Bioinforma.*, vol. 42, no. 1, p. 3.1.1-3.1.8, Jun. 2013, doi: 10.1002/0471250953.bi0301s42.

[33] A. M. Schnoes, S. D. Brown, I. Dodevski, and P. C. Babbitt, “Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies,” *Plos Comput Biol*, vol. 5, 2009.

[34] M. L. Green and P. D. Karp, “Using genome-context data to identify specific types of functional associations in pathway/genome databases,” *Bioinformatics*, vol. 23, pp. 205–211, 2007.

[35] Y. Matsuta, M. Ito, and Y. Tohsato, “ECOH: An Enzyme Commission number predictor using mutual information and a support vector machine,” *Bioinformatics*, vol. 29, pp. 365–372, 2013.

[36] M. Y. Galperin and E. V. Koonin, “Divergence and Convergence in Enzyme Evolution,” *J. Biol. Chem.*, vol. 287, no. 1, pp. 21–28, Jan. 2012, doi: 10.1074/jbc.R111.241976.

[37] Y. Ofran and H. Margalit, “Proteins of the same fold and unrelated sequences have similar amino acid composition,” *Proteins Struct. Funct. Bioinforma.*, vol. 64, no. 1, pp. 275–279, Jul. 2006, doi: 10.1002/prot.20964.

[38] V. Giri, T. V. Sivakumar, K. M. Cho, T. Y. Kim, and A. Bhaduri, “RxnSim: a tool to compare biochemical reactions,” *Bioinformatics*, vol. 31, pp. 3712–3714, 2015.

[39] Q. N. Hu, Z. Deng, H. A. Hu, D. S. Cao, and Y. Z. Liang, “RxnFinder: biochemical reaction search engines using molecular structures, molecular fragments and reaction similarity,” *Bioinformatics*, vol. 27, pp. 2465–2467, 2011.

[40] Y. Moriya, T. Yamada, S. Okuda, and Z. Nakagawa, “Identification of Enzyme Genes Using Chemical Structure Alignments of Substrate-Product Pairs,” *J. Chem. Inf. Model.*, vol. 56, pp. 510–516, 2016.

[41] Q. N. Hu, H. Zhu, X. B. Li, and M. M. Zhang, “Assignment of EC Numbers to Enzymatic Reactions with Reaction Difference Fingerprints,” *Plos One*, p. 7, 2012.

[42] S. A. Rahman, S. M. Cuesta, N. Furnham, G. L. Holliday, and J. M. Thornton, “EC-BLAST: a tool to automatically search and compare enzyme reactions,” *Nat. Methods*, vol. 11, no. 2, pp. 171–174, Feb. 2014, doi: 10.1038/nmeth.2803.

[43] *DAYLIGHT, Version 4.62, DAYLIGHT Inc., Mission Viejo, CA.* .

[44] D. J. Rogers and T. T. Tanimoto, “A Computer Program for Classifying Plants,” *Science*, vol. 1960, no. 132, pp. 1115–1118.

[45] International Union of Biochemistry and Molecular Biology and E. C. Webb, Eds., *Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes*. San Diego: Published for the International Union of Biochemistry and Molecular Biology by Academic Press, 1992.

[46] N. Hadadi, J. Hafner, A. Shajkofci, A. Zisaki, and V. Hatzimanikatis, “ATLAS of Biochemistry: A Repository of All Possible Biochemical Reactions for Synthetic Biology and Metabolic Engineering Studies,” *ACS Synth. Biol.*, vol. 5, no. 10, pp. 1155–1166, Oct. 2016, doi: 10.1021/acssynbio.6b00054.

[47] N. Hadadi, J. Hafner, K. C. Soh, and V. Hatzimanikatis, “Reconstruction of biological pathways and metabolic networks from in silico labeled metabolites,” *Biotechnol. J.*, vol. 12, no. 1, p. 1600464, Jan. 2017, doi: 10.1002/biot.201600464.

[48] H. Briem and U. F. Lessel, “In vitro and in silico affinity fingerprints: Finding similarities beyond structural classes,” in *Virtual Screening: An Alternative or Complement to High Throughput Screening?*, vol. 20, G. Klebe, Ed. Dordrecht: Kluwer Academic Publishers, 2002, pp. 231–244.

[49] N. M. O’Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, “Open Babel: An open chemical toolbox,” *J. Cheminformatics*, vol. 3, no. 1, p. 33, 2011, doi: 10.1186/1758-2946-3-33.

[50] D. Weininger, “SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules,” *J. Chem. Inf. Comput. Sci.*, vol. 28, no. 1, pp. 31–36, Feb. 1988, doi: 10.1021/ci00057a005.

[51] L. Leydesdorff, “On the normalization and visualization of author co-citation data: Salton’s Cosineversus the Jaccard index,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 59, no. 1, pp. 77–85, Jan. 2008, doi: 10.1002/asi.20732.

[52] P. Carbonell *et al.*, “Selenzyme: enzyme selection tool for pathway design,” *Bioinforma. Oxf. Engl.*, vol. 34, no. 12, pp. 2153–2154, Jun. 2018, doi: 10.1093/bioinformatics/bty065.

[53] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic Local Alignment Search Tool,” *J. Mol. Biol.*, vol. 215, pp. 403–410, 1990.

[54] A. Burgard, M. J. Burk, R. Osterhout, S. Van Dien, and H. Yim, “Development of a commercial scale process for production of 1,4-butanediol from sugar,” *Curr Opin Biotechnol*, vol. 42, pp. 118–125, 2016.

[55] S. Andreozzi, A. Chakrabarti, K. C. Soh, and A. Burgard, “Identification of metabolic engineering targets for the enhancement of 1,4-butanediol production in recombinant E. coli using large-scale kinetic models,” *Metab. Eng.*, vol. 35, pp. 148–159, 2016.

Chapter 5 Enzyme prediction in practice: lessons learned, challenges and opportunities

"We cannot solve our problems with the same level of thinking that created them"
Albert Einstein

5.1 Introduction

A part of section 5.1 is submitted as a chapter for metabolic engineering series books (pathway design). Dr. Jasmin Hafner and the author both were in charge of preparation and writing.

As the field of biotechnology, and in particular metabolic engineering and synthetic biology grows, applications increasingly turn toward the production of valuable industrial chemicals and new methods for designing de novo synthetic pathways. The biosynthesis of added value compounds offers several key advantages over chemical synthesis. First, various enzymes are required to catalyse different steps of a complex biosynthesis pathway, can operate simultaneously under the same biological relevant conditions, such as: ambient temperature and pressure, neutral pH and aqueous solution. In addition, the renewable and cheap materials can be used as the input of process and finally higher amount of product can be achieved by consumption of less amount of energy [1] .

In the biosynthesis, the compound to be produced is called "target compound". The target compound is synthesized from one (or several) starting metabolites, called the "precursor compound". The biosynthetic pathway is therefore defined as a sequence of reactions that convert the precursor compound into the final target compound. Metabolic pathways can be designed *manually* by relying on the intuition of the scientist or be generated by computational predictive tools. It is up to the scientist to decide which part of the design step can be done manually, and which one needs a computational approach. In general, if the pathway leading to the target compound has been characterized previously in another organisms, the easiest solution might be to rely on existing knowledge and to express the enzymes catalyzing each step in the host (heterologous pathway expression). For this, the scientist can consult biochemical databases to retrieve information on the pathway to be implemented. The main drawback of manual pathway design is that the

pathway chosen by intuition might be suboptimal compared to other unexplored biochemical possibilities. Also, this approach is limited to known biochemistry from scientific articles and databases, and not considering novel predicted reactions. The success of manual pathway design relies on the biochemical knowledge of the researcher. In the beginning of metabolic engineering, intuitive pathway design has been the only possible approach, and since then this method is behind many success stories of metabolic engineering. Later, advanced computational tools and methods accelerated the development of bioproducer strains and stimulated the exploration of the biochemical space.

Whether a pathway comes from a known database or has been generated by computational tools, each enzymatic reaction steps in the pathway needs to be catalyzed by an enzyme. For known, well-described reactions, appropriate enzymes can be found by literature search or database lookup. If the list of pathways obtained from the previous step is long, automated querying of enzyme databases can be used to assign enzymes to reactions. However, the enzyme information cannot be found if the pathway incorporates orphan reactions (i.e, enzyme catalyzing the reaction is not discovered yet, see chapter 4) or novel reactions (reactions predicted by predictive computational tools, see chapter 3).

Despite the increasing progress in enzyme discovery, still about one fourth of the enzymatic reactions catalogued in biochemical databases are orphan, meaning these observed enzymatic activities are not associated with gene information in any organism [2]. This type of orphan reactions is defined as global orphan. From taxonomic point of view, a second type of orphan activity, known as local orphan is defined [2]. Local orphan reactions are observed enzymatic activities in at least one organism of a clade although the corresponding gene information only exist in other clades [2]. An example is the aspartate 4-decarboxylase (EC 4.1.1.12) which catalyses the conversion of L-aspartate to L-alanine by releasing carbon dioxide. According to uniport, EC 4.1.1.12 is linked to hundreds of gene sequence in bacteria. However, there is not any sequence annotated with this function in eukaryote or archaea kingdom. Nevertheless, the activity of aspartate 4-decarboxylase is observed and characterized in several mammalians. Therefore, EC 4.1.1.12 is a local orphan activity in eukaryotes. In case of archaea, there is no literature evidence or characterization for activity of this enzyme reported so far, consequently this enzymatic activity is considered absent in this kingdom [2]. In conclusion, in addition to one fourth of global orphan activities, the portion of local orphan reactions in each kingdom is significant and changes from 24 percent in archaea to 20 percent in eukaryote and 14 percent in bacteria [2] and leaves important gaps in metabolic pathways.

Furthermore, producing new-to-nature compounds (e.g., biosynthesis of chemicals) necessarily entails the engagement of novel reactions in the pathway and therefore use of novel enzymatic activities. But the question is, “which enzymes can generate such metabolic novelties?”. Most if not all the enzymes are promiscuous, meaning they can catalyze side reactions other than their main function. In nature, the presence of promiscuity activities has led to evolution of new enzymes [3]. However, the promiscuity and

basically desired non-native function of an enzyme will not be known *a priori*. Given the wealth of enzymatic knowledge that has been accumulated, a computational method to predict enzymes that may catalyze a desired transformation will greatly expedite the development of biosynthetic pathways engineered to produce new-to-nature products. Understanding the relation between structure and promiscuity remains a challenging feature to be understood. Enzyme prediction tools such as BridgIT[4], EC-BLAST[5], E-zyme[6] and Selenzyme[7] determine the structural similarity of a novel reaction to all well-characterized reactions in biochemical databases, and propose a list of enzyme candidates ranked by their likelihood to catalyze the desired transformation. The novelty of BridgIT tool relies on using reactive site-centric similarity, meaning instead of considering the overall structure of molecules in similarity calculations, BridgIT calculates similarity in reactive site and the neighborhood around the reactive site (Please see chapter 4). In contrast, the overall structure can be much larger than the reactive site and skew the comparison by indicating high similarities when the reactivity is actually quite different [8].

In this chapter, we demonstrate capability of BridgIT in annotation of orphan reactions via two case studies. In the first case study (section 5.2), we aim to predict a homologous pathway for adipic acid biosynthesis in yeasts and specifically in *Saccharomyces cerevisiae* and *Yarrowia lipolytica*. Adipic acid is a chemical building block for the production of nylon and polyurethane. Adipic acid is currently produced by catalytic oxidation of benzene derivatives with concurrent production of nitrous oxide, which contributes to the greenhouse effect and ozone layer depletion. A more environmentally friendly bio-based production process is desirable. To design a route towards adipic acid, a conventional literature and database search was performed. Finally, an existing pathway with 10 steps of KEGG reactions for the production of adipic acid proposed. However, 4 steps (of 10) were the bottlenecks for the production, since 2 were global orphan reactions and the other two reactions were local orphan. Using BridgIT, we discovered 4 enzymes (native to *Yarrowia lipolytica)* for catalyzing each orphan step and we demonstrated its functionality in the yeasts *Saccharomyces cerevisiae* and *Yarrowia lipolytica*. *Y.lipolytica* was engineered by over-expression of homocitrate synthase *YALI0F31075g* (E.C.2.3.3.14), homoaconitate hydratase *YALI0E02728g* (E.C.4.2.1.36), di- and tri-carboxylic acids mitochondrial transporters *YALI0D02629g* and *YALI0F26323g*, and by expression of codon-optimized adipate-semialdehyde dehydrogenases from *Acinetobacter* sp. and *Pseudomonas* sp. The engineered strain produced 0.2 mg/L of adipic acid in mineral medium with glucose as the sole carbon source and 30 mg/L adipic acid in municipal solid waste hydrolyzate. The work demonstrates the utility of BridgIT for pathway discovery and describes the first biosynthetic route towards adipic acid that functions in eukaryotes.

In the second case study, the plant specialized metabolism as an important source of pharmaceutical molecules is studied. Despite their significant value only a fraction of plant natural products (PNPs) and their derivatives have been explored. To access this untapped potential, the reconstitution of heterologous PNP biosynthesis pathways in engineered microbes provides a valuable starting point to explore and produce

novel PNP derivatives. We introduce a computational workflow to systematically screen the biochemical vicinity of a biosynthetic pathway for pharmaceutical compounds that could be produced by derivatizing intermediates in the original pathway. As a case study, we explored the neighborhood of noscapine pathway, a benzylisoquinoline alkaloid with a long history of medicinal use. We found (S)-tetrahydropalmatine, a known analgesic and anxiolytic, is one reaction step away from intermediates of noscapine pathway. However, the last step is a global orphan reaction. We used BridgIT to find candidate enzymes for its catalyzation. The two-top proposed candidates exhibited the desired activity, resulting in a yeast platform for (S)-tetrahydropalmatine production. Our novel approach provides a valuable resource for researchers who aim to study and engineer the bioproduction of natural product derivatives.

These case studies demonstrate the value of cheminformatic tools to predict reactions, pathways, and enzymes in synthetic biology and metabolic engineering. Furthermore, studying these specific situations offer great opportunities to investigate the variety of challenges that the biosynthesis and pathway design could face in practice. The focus is not so much on the specific compounds to be produced but on the type of problems, how they are understood and approached, and what action ensues. In conclusion, the findings from these two case studies offer guidance toward (i) manual or computational design of pathways, (ii) homologues or heterologous pathway expression and their implementation, (iii) local or global orphan enzymatic activities and their annotation (Table 5.1).

Table 5.1: Overview on the pathways discussed in section 5.2 and 5.3.

Target Compound	Organism	Pathway design	Orphan reactions		Pathway type*	Highest titer	Carbon source
			Global	Local			
Adipic acid	*Y. lipolytica* *S. cerevisiae*	Manual	**2****	**2****	Homologous	30 mg/L	Food waste hydrolysate
Tetrahydropalmatine (THP)	*S. cerevisiae*	Computational	**1****	-	Heterologous	3.45 µg/L	Sugar

* Homologous or heterologous. Homologous pathways are catalyzed by native enzymes of the organisms. In case of heterologous pathways, additional enzymes are integrated into microorganism.
** BridgIT is used as the enzyme annotation tool.

Due to the high applicability of BridgIT for any metabolic or protein engineering application, the proposed candidate enzymes will be an important resource for all academic and industry researchers actively involved in protein evolution and engineering technologies for nearly any biotechnology application. It is reasonable to expect that BridgIT method, with its wide – reaching multidisciplinary aspect, will serve as a tool for categorizing future enzymatic reactions and as a reference database for all researchers for the next generations of enzymatic technology development.

5.2 A novel pathway for adipic acid biosynthesis in yeasts

This subchapter is the result of collaboration with the experimental lab of Prof. Irina Borodina at University of Denmark (DTU), which started in the context of two weeks exchange student part of PAcMEN EU project (Predictive and Accelerated Metabolic Engineering Network) [2]*. The results of this collaboration led to discovery of a new biosynthetic pathway for production of adipic acid in yeast. This subchapter represents the manuscript corresponding to this study which is recently submitted. In this project, the experimental results and BLAST analysis have been obtained by Ksenia Chekina, Dmitriy Abashkin, Jonathan Dahlin, Mette Kristensen, Nicholas Milne and Maria Sanchis. BridgIT analysis and pathway feasibility study have been provided by the author of this thesis. Prof. Vassily Hatzimanikatis and Prof. Irina Borodina supervised the project as well as the completion of the manuscript.*

Full list of authors of this paper: K. Chekina†, H. MohammadiPeyhani†, D. Abashkin, J. Dahlin, M. Kristensen, N. Milne, M. Sanchis, V. Hatzimanikatis, and Irina Borodina*," A novel pathway for adipic acid biosynthesis in yeasts" († Contributed equally, * corresponding authors).*

5.2.1 Introduction

Adipic acid (hexane-1,6-dioic acid) is a dicarboxylic acid with a $5.6 billion market in 2016 [9]. Adipic acid is primarily used for the production of nylon-6,6 in a polycondensation reaction with hexamethylenediamine. Other uses include production of other plastics and application as an acidity regulator in foods.

Adipic acid is made through oxidation of cyclohexanone and cyclohexanol with nitric acid, where a by-product nitrous oxide (N_2O) is produced in a one-to-one molar ratio to adipic acid [10]. Nitrous oxide is a potent greenhouse gas with a 265–298 times higher global warming potential than CO_2[11]. It is also currently the main ozone-depleting chemical[12]. Recently, catalytic methods have been developed to reduce the emissions of N_2O in the tail gasses from adipic acid production by up to 98%[13]. However, the remaining 2% still released is equal to 4 million tons of CO_2 per year and remains to be the second-largest source of industrial N_2O pollution[14]. The feedstocks for adipic acid production are crude oil and naphtha. There are hence ongoing efforts to develop novel processes for adipic acid production from renewable feedstocks.

The first attempt to produce adipic acid from biomass was reported in the 1980s. A two-step process was demonstrated, where 1,6-hexanediol was generated by dehydration and hydrogenation of lignocellulosic biomass, and, consequently, 1,6-hexanediol was oxidized into adipic acid biologically using *Gluconobac teroxydans* subsp. *oxydans*[15]. The process was, however, very energy-intensive and was not realized at scale.

Another possible route is a two-stage process, where in the first step, *cis,cis*-muconic acid (CCM) is produced from lignin or sugars, and in the second step, CCM is hydrogenated into adipic acid via chemical catalysis[16]. The whole process from lignin to nylon was demonstrated at laboratory scale, where hydrothermally depolymerized softwood lignin was used as the substrate for CCM production by an engineered strain of *P.*

[2] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 722287.

putida, resulting in 13 g/L CCM titer. CCM was purified from the fermentation broth, hydrogenated into adipic acid, which in turn was polymerized into nylon-6,6.[17]

Several studies reported the production of CCM from glucose, with the highest titer for *E. coli* being 16 g/L[18], and for the yeast *S. cerevisiae* being 20.8 g/L CCM in controlled fed-batch fermentation, with a yield of 8.4% ($mol_{CCM}/mol_{glucose}$) [19].

Direct production of adipic acid through bioconversion of fatty acids by engineered oleaginous yeast *Candida tropicalis* was developed by the company Verdezyne. The strain expressed heterologous ω-oxidases and had some peroxisomal oxidases deleted to ensure that fatty acyl-CoAs were preferentially truncated down to six carbons to give adipic acid[20]. The disadvantages of the process were a relatively high cost of the fatty acid feedstocks and low yields due to carbon loss in the β-oxidation.

The only organism so far reported to naturally produce adipic acid is a thermophilic bacterium *Thermobifida fusca* of the order *Actinomycetales* [21]. The pathway was elucidated to be comprised of five steps, (i) ligation of succinyl-CoA and acetyl-CoA into 3-oxoadipyl-CoA, (ii) reduction into 3-hydroxyadipyl-CoA, (iii) dehydration to obtain a double-bond, and (iv) hydrogenation into adipyl-CoA, which is (v) lyased to give adipic acid. Overexpression of 5-carboxy-2-pentenoyl-CoA reductase (Tfu_1647) in *T. fusca* B6 resulted in 2.23 g/L adipic acid produced from glucose as the sole carbon source with 5.6% yield ($mol_{adipate}/mol_{glucose}$).

The pathway from *T. fusca* was expressed in *E. coli*, giving 2.5 g/L of adipic acid in bioreactors using glycerol as a carbon source, with 4.4% yield ($mol_{adipate}/mol_{glycerol}$)[22]. To our knowledge, this pathway has not yet been successfully expressed in eukaryotic hosts, likely due to the problems with the expression of a functional iron-sulfur cluster-dependent dehydratase.

Here we investigate the possibility to produce adipic acid from sugars using yeasts. Several yeast species are particularly well suited for large-scale industrial processes, due to their resistance to the stresses in large tanks, low pH tolerance, and long history of safe use, among them the oleaginous yeast *Yarrowia lipolytica* and baker's yeast *S. cerevisiae*.

5.2.2 Materials and methods

5.2.2.1 Computational pathway design

To design a route towards adipic acid, a conventional literature and database search was performed. Then, the possible targets were analyzed by BLAST[23] and BridgIT[24] algorithms and finally were imbedded into a *Yarrowia* genome-scale model, iYali4[25], for feasibility analysis. The updated model was used for flux balance analysis (FBA)[26] and thermodynamic based flux analysis (TFA)[27] to interrogate the mass balance and energy balances of the pathways respectively.

The BLASTp algorithm with the default parameters was used to find endogenous alternatives in *Y. lipolytica* to thermophilic enzymes. CLC Main Workbench 8 (https://digitalinsights.qiagen.com/) was used to align Lys21p of *S. cerevisiae* to the YALI0F31075p of *Y. lipolytica*.

We used the BridgIT algorithm, to find the potential promiscuous enzymes able to catalyze orphan reactions inside pathway. In this study, we used the online version of BridgIT (https://lcsb-databases.epfl.ch/Bridgit) with default parameters as suggested by the original method [24]. We considered a BridgIT score of 0.3 as the minimum standard threshold in similarity evaluations.

Furthermore, the genome scale model of *Y. lipolytica*, iYali4, which has 1942 reactions and 1691 metabolites was used for flux balance analysis (FBA)[26] and thermodynamic based flux analysis (TFA)[27] to interrogate the mass balance and energy balances of the pathways respectively. In FBA feasibility is analysed using stoichiometric mass balance constraints and default reaction directionalities in the model [26]. In TFA, thermodynamic constraints are used to assign reaction directionalities that are thermodynamically favorable. These constraints take into account Gibbs free energy of reactions and metabolite concentrations[27].

5.2.2.2 Microorganisms

Chemically competent *Escherichia coli* DHα was used for the cloning and plasmid propagation. Transformed with plasmids, *E. coli* cultures were grown in Lysogeny Broth (LB) supplemented with 100 mg/L ampicillin and when needed with 15 g/L agar.

Two strains of *Yarrowia lipolytica* and one strain of *Saccharomyces cerevisia*e were subject to genome editing. *Y. lipolytica* Y-63746 (MatA, *Y. lipolytica* W29) was a kind gift from ARS Culture Collection, NCAUR, USA. *Y. lipolytica* GB20 with genotype MATb, *ku70Δ, nugm-Htg2, ndh2i, lys11⁻, leu2⁻, ura3⁻* was a kind gift of Volker Zickermann (Goethe-Universität, Germany). *S. cerevisiae* CEN.PK113-7D, a prototrophic haploid strain (*MATa URA3 HIS3 LEU2 TRP1 MAL2-8^{c} SUC2*) was a gift from Dr. Peter Kötter (Goethe-Universität, Germany).

Y. lipolytica strains were edited according to EasyCloneYALI protocol[28]. *S. cerevisiae* strains were modified according to EasyClone-MarkerFree protocol[29].

The strains are described in appendix Table 8.4, plasmids in Table 8.5, biobricks in Table 8.6, and primers in appendix Table 8.7. Heterologous genes were codon-optimized for *Y. lipolytica* and ordered as synthetic gene strings from GeneArt, Thermofisher (appendix Table 8.8).

5.2.2.3 Media and cultivation conditions

All strains were stored as cryostocks, at -80°C in medium with 30% v/v glycerol. The composition of mineral medium for *S. cerevisiae* was as described previously[30], with major components being 7.5 g/L (NH4)2SO4, 14.4 g/L KH2PO4, 0.5 g/L MgSO47H2O, 20 g/L glucose, 1 mL of trace metals solution, and 1 mL vitamins. In some experiments, the medium was supplemented with 0.2 g/L of 2-oxoglutarate. For *Y. lipolytica,* the

medium was the same, except that the trace metal mix was as specified by Kamzolova et al.[31]. For growing auxotrophic strains, the medium was supplemented with 76 mg/L lysine, 380 mg/L leucine, and 20 mg/L uracil. All the media was adjusted to pH 4.5 and filter-sterilized through bottle top filters with 0.2 µm pore size.

The food waste hydrolysate was prepared from a batch of canteen leftovers. It was freeze-dried and milled using Polymix PX-MFC mill, Kinematica (WH1, WH2, WH3-1, WH3-3) or just wet-milled (WH4-2, WH4-3) followed by total solids adjustment to 25%. Then WH4-3 sample were treated with commercial proteases TRIO 15 g/100 g glucane at 50°C for 8 hours with inactivation of proteases at 80°C for 2 hours. Then all samples were adjusted to pH 5.2, and commercial cellulases Cellic Ctec3, 15 g/100 g glucane, were applied for 72 hours at 50°C. Then WH1, WH2, and WH3-3 samples were directly frozen, WH3-1 was supplemented with 4 mg/L of ampicillin and then frozen, and WH4-2, and WH4-3 samples were autoclaved at 121°C for 1 hour before freezing.

The samples of the pre-treated hydrolysate were thawed and centrifuged in 50 mL Falcon tubes at the 15,000 g for 10 minutes. After removing the insoluble fraction with a cotton cloth, the supernatant was filter-sterilized through syringe filters with 0.2 µm pores, and directly used for further experiments. 200 µL of each sample was submitted for HPLC analysis for the sugar concentration measurements (UltiMate 3000, Dionex) using Aminex HPX-87H ion exclusion column with a 5 mM H_2SO_4 flow of 0.6 ml/min for 45 min per sample and the temperature of column 50°C. Sugars were detected using RI-101 Refractive Index Detector (Dionex). The data were acquired and analyzed with Chromeleon software using the correlation curve of the standards with known concentrations.

For cultivation test, 2 ml of yeast peptone dextrose medium (YPD, Sigma–Aldrich) in 13 ml-round bottom tubes were inoculated directly from the cryostocks and incubated overnight at 30°C with 250 rpm agitation. The cells were harvested by centrifugation in 2 ml sterile Eppendorf tubes at 3,000 x g for 5 min, washed twice with sterile water, and diluted in water until OD 10, controlled by a spectrophotometer (NanoPhotometer Perl, Implen Gmbh, Munich, Germany).

The 24-well clear-bottom plates (Porvair Sciences, Leatherhead, UK) containing 1 ml of medium (2% glucose mineral medium or food waste hydrolysate) were inoculated with pre-washed culture with starting OD 0.05 and incubated at 30°C for 70h with 250 rpm at 100% humidity in the Growth Profiler 960 (Enzyscreen, Haarlem, The Netherlands).

The endpoint samples were diluted with water with ration 1/3, filter sterilized and submitted for LS-MS analysis.

The growth data was collected by processing the phase-contrast images acquired with 809 magnification every 15 minutes, converting the G-value into OD using the OD600 equivalent = 0,01 × (GV-GB)^1,5807, where GV is a G-value of the well, and GB is a G value of the plain media. The correlation was found by measuring several growth points at the photometer and refereeing them to G-value of the taken pictures.

5.2.2.4 Metabolite measurement by LS-MS

The concentration of adipic acid and the pathway intermediates in the broth was measured on LC-MS system, Dionex UltiMate 3000 UHPLC (Fisher Scientific, San Jose, CA) connected to an Orbitrap Fusion Mass Spectrometer (Thermo Fisher Scientific, San Jose, CA). The system used a Waters ACQUITY HSS T3 C18 UHPLC column, with a 1.8 um particle size, 2.1 mm i.d. and 100 mm long kept at 30°C. The flow rate was 0.400 mL/min with 0.1% formic acid (A) and 0.1% formic acid in acetonitrile (B) as the mobile phase. The gradient started at 5% B for 1.5 min and then followed a linear gradient to 60% B over 5 min. This solvent composition was held to 5.5 min after which it was changed immediately to 90% B until 6.0 min. Finally, the gradient was changed to 5% B until 8 min. The sample (2 uL) was passed on to the MS equipped with a heated electrospray ionization source (HESI) with sheath gas set to 45 (a.u.), aux gas to 13 (a.u.) and sweep gas to 1 (a.u.). The cone and probe temperatures were 342°C and 358°C, respectively. Spray voltage was 2500 V in negative ionnization mode and 3500 V in positive mode. Scan range was 100 to 600 Da. Detection of adipic acid (145.0506 ion),2-oxoadipate (159.0299 ion), 2-aminoadipate (160.0615 ion),2-semialdehyde (129.0557 ion) and 2-oxopimelate (173.0444 ion) was conducted in full scan. Quantification of 2-oxoadipate, 2-aminoadipate, and adipate was based on calculations from calibration standards analyzed before and after sets of 32 samples. The concentrations of 2-oxopimelate and adipate-semialdehyde in the broth of different strains were compared by using relative area units of ions with masses matching the theoretical m/z of the specific compounds.

All reagents used were of analytical grade and purchased from Sigma-Aldrich, except adipic acid, which was purchased from TCI Europe N.V.

5.2.3 Results

5.2.3.1 Pathway design

The adipic acid pathway previously reported in the bacterium *T. fusca* has not yet successfully been implemented in yeast. The focus of the current study was to find an alternative route for *de novo* synthesis of adipic acid in yeast that does not require Fe-S cluster-dependent enzymes.

Adipic acid is a dicarboxylic acid with 6 carbons. Among native compounds in yeast, the closest compound is 2-oxoadipic acid, which is an intermediate of lysine biosynthesis. During lysine biosynthesis, homocitrate synthase combines the central carbon metabolites 2-oxoglutarate and acetyl-CoA to make homocitrate. Homocitrate is transported into mitochondria, where it is isomerized into homoisocitrate via homoaconitate.

Homoisocitrate, in turn, is converted into 2-oxoadipate by homoisocitrate dehydrogenase. 2-oxoadipate is exported into the cytosol and transaminated to give the lysine precursor 2-aminoadipate. We hypothesized that if homocitrate synthase could accept 2-oxoadipate as the substrate instead of 2-oxoglutarate and the resulting one-carbon longer compound would go through the same steps, then 2-oxopimelate would be produced. The 2-oxopimelate could be decarboxylated to give adipate semi-aldehyde, which could be further oxidized into adipic acid.

The majority of decarboxylation and oxidation enzymes in nature are known to be active on a wide range of substrates[32],[33],[34]. The promiscuity of lysine biosynthetic enzymes has been described in methanogenic bacteria. Methanogen homoaconitase (E.C.4.2.1.114) participates in the chain extension in methanogenic bacteria in the order: 2-oxoglutarate – 2-oxoadipate – 2-oxopimelate – 2-oxosuberate. The extension of 2-oxoglutarate to 2-oxoadipate in methanogens proceeds via the same intermediates as in yeast and the reaction mechanisms of every next loop are similar to those from the first extension. Homoaconitase of methanogens, such as *Methanocaldococcus vulcanius*, *Methanothermobacter thermautotrophicus*, and many others, consists of two subunits *aksD* and *aksE*, which require Fe-S cluster[35]. We therefore decided to find alternatives.

The BLASTp algorithm with default settings was used to align the amino-acid sequence of the large and small subunits of all 121 annotated in KEGG metanogen homoaconitases to *Y. lypolytica* genome (CLIB122) to investigate if *Y. lipolytica* carries similar enzymes (appendix Table 8.9). The homoaconitase YALI0E02778p came up as the closest homologue with identity of 29-38% for the large subunit and 29-52% for the small subunit. YALI0E02778p homoaconitase is responsible for the isomerisation of homocitrate into homoisocitrate via dehydratation-hydratation steps (Figure 5.1).

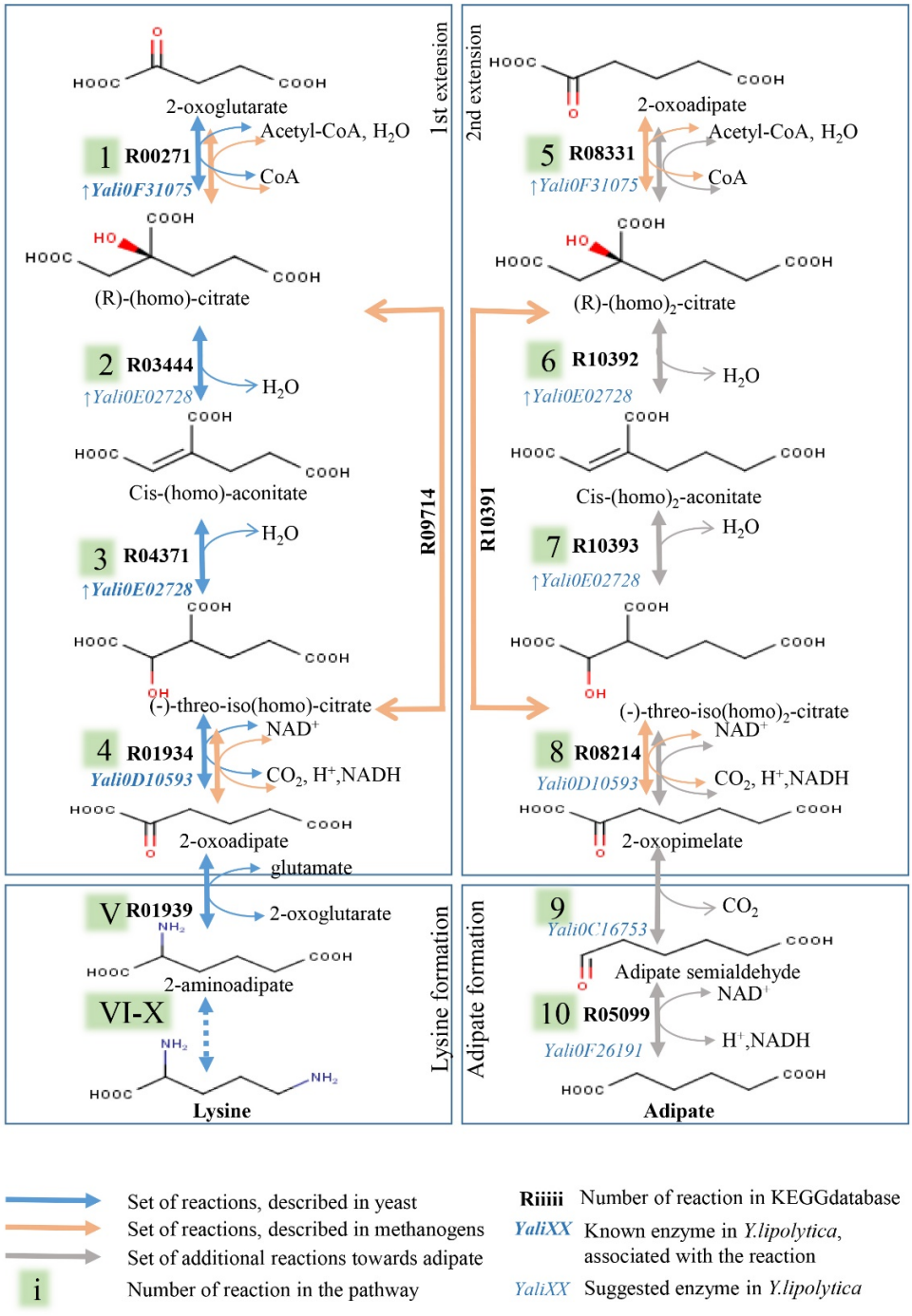


Figure 5.1: Proposed adipic acid biosynthetic pathway. Steps 1 to 10 indicate the pathway towards adipic acid, which brunches from lysine biosynthesis with is steps 1 to 4 and V to X combined. Yali genes ID indicate described (in bold) and suggested (normal) genes in *Y. lipolytica* genome, and ↑ stands for overexpression of the genes in the current study.

Next, we applied BridgIT algorithm to predict whether the enzyme would be able to isomerize $(homo)_2$-citrate into $iso(homo)_2$-citrate as well. BridgIT is a computational tool that identifies candidate enzymes for an input reaction based on the promiscuity of enzymes and introduces the information of the enzyme binding pocket

into reaction similarity comparisons [24]. It ascertains the similarity of two reactions by comparing the reactive sites of their substrates and neighborhood of reactive sites, along with the structures of the generated products. BridgIT compares orphan and novel reactions to enzymatic reactions with known protein sequences, and then, it proposes protein sequences and genes of the most similar non-orphan reactions as candidates for catalyzing the novel or orphan reactions. BridgIT suggests a list of top-ranked candidate EC numbers for every orphan/novel reaction, which allows us to evaluate other possible EC numbers even if the first candidate is not associated with any gene in the target organism. In this study, we used BridgIT tool to find (1) candidate protein sequence for orphan reactions (R08214, and R08331) and (2) alternative native candidate enzymes for non-native reactions in *Y. lipolytica* (R10392 and R10393, native to *Methanocaldococcus jannaschii*). Figure 5.2 demonstrates BridgIT procedure for an orphan reaction "R08214" (Figure 5.2). We followed the same procedure for the remaining 3 reactions, the result of the most similar reactions along with their EC numbers are shown in Table 5.2.

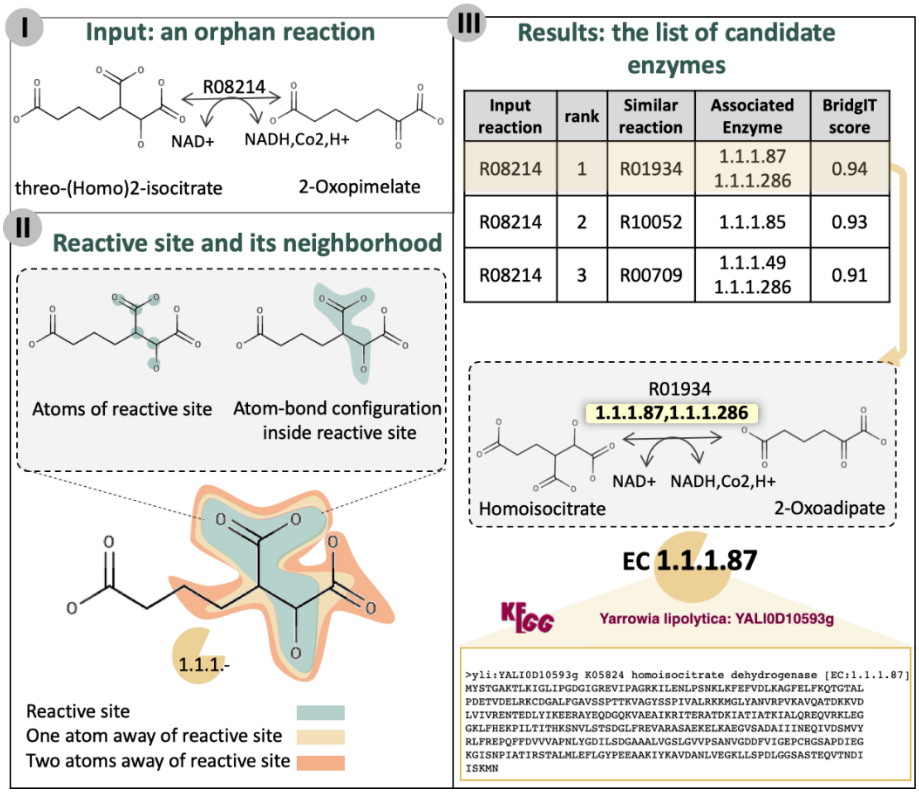


Figure 5.2: BrigIT workflow proposes promiscuous enzymes for an orphan reaction. (I) Input to the workflow is an orphan reaction, R08214, which decarboxylates threo-(Homo)2-isocitrate. (II) BridgIT scans the substrates with enzymatic reaction rules and identifies the reactive site (green shade). The information about reactive site and its neighborhood (until seven atoms away from reactive site) along with corresponding atoms on the products are used for similarity evaluation. (III) The result report, ranks the most similar non-orphan reactions based on their similarity to input. Here, R01934 catalyzed by EC 1.1.1.87 or EC1.1.1286 is the top ranked with BridgIT score 0.94. EC 1.1.1.87 is annotated as YALI0D10593g in *Y. lipolytica* genome.

Table 5.2: BridgIT results specific to *Yarrowia lipolytica* for R08331, R10392, and R10393 (only top results are shown).

Input reaction	EC Associated Enzyme	Similar KEGG reaction	EC Associated Enzyme to similar reaction	Gene	BridgIt score
R08331	orphan	R08640	EC 2.3.3.14	YALI0F31075g	0.77
R10392	EC 4.2.1.114	R04371	EC 4.2.1.36	YALI0E02728g	0.74
R10393	EC 4.2.1.114	R04371	EC 4.2.1.36	YALI0E02728g	0.72

The pathway (10 reactions, Figure 5.1) was embedded in the curated *Y. lipolytica* genome-scale model (iYali4[25]), and then FBA and TFA feasibility were checked. Conversion of cis-$(homo)_2$-aconitate to $(homo)_2$-isocitrate (R10393) is thermodynamically not feasible on its own. In methanogens, this reaction is coupled with the previous reaction (R10392) under the control of one enzyme channeling the two reactions in one two-step reaction (R10391)[35]. When considered as such, the whole pathway is mass-balanced and energetically feasible in the context of the *Y. lipolytica* metabolic network with a maximum theoretical yield of 0.567 g adipid acid per g glucose consumed.

5.2.3.2 Production of adipic acid via the lysine pathway in *Yarrowia lipolytica*

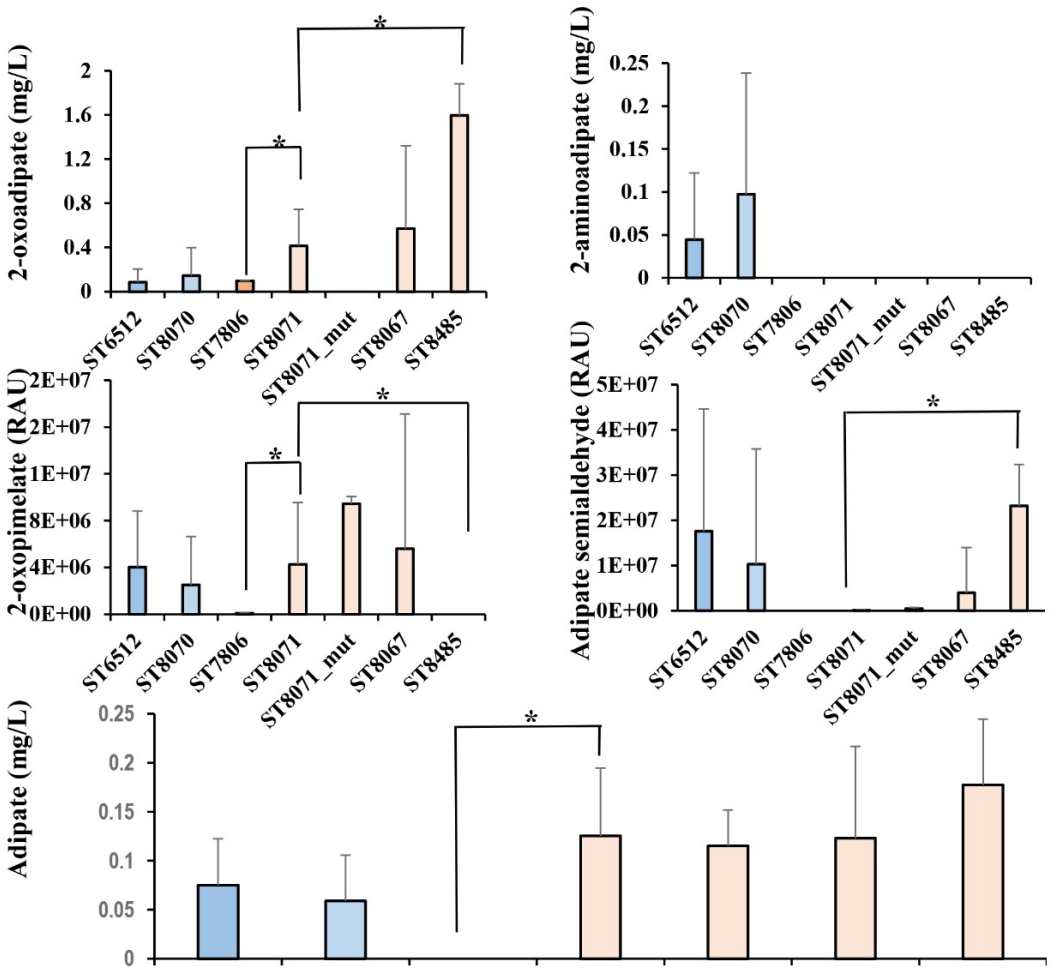
The computational analysis suggested that adipic acid can be produced in *Y. lipolytica* with the native enzymes. Indeed, a commonly used laboratory *Y. lipolytica* strain W29 produced 75±42 μg/L of extracellular adipic acid when cultivated in the mineral medium with glucose as the sole carbon source. The titer was very low, but detectable by LC-MS. (Appendix, Figure 8.2).

To validate that adipic acid was produced as a side product of lysine biosynthesis, we cultivated another non-engineered laboratory strain of *Y. lipolytica* GB20, which had a defective homocitrate synthase gene (*YALI0F31075g*)[36]. The strain was cultivated on the same medium, but with supplementation of lysine, leucine, and uracil to compensate for auxotrophy. No adipic acid was detected in the broth. We detected 96±1.5 μg/L 2-oxoadipate in the broth, but this can likely be attributed to lysine degradation. Next, we tested whether restoration of homocitrate synthase activity would enable adipic acid production. Expression cassettes for *YALI0F31075g* (homocitrate synthase) and *YALI0E02728g* (homoaconitate hydratase) were integrated into a GB20 strain, and the resulting strain produced 125±69 μg/L of adipic acid. 2-oxoadipate in concentration 412 μg/L on average with high variation (standard deviation was±332 μg/L) and some 2-oxopimelate was potentially detected, though the latter seven-carbon intermediate was not found in the parental strain. The exact concentration of 2-oxopimalate was not possible to calculate due to the lack of

standard, but the ion-peak matching the accurate mass of 2-oxopimelate was detected (Appendix, Figure 8.2).

The overexpression of the same genes in W29 background increased 2-oxoadipate up to 144±22.5 μg/L compared to 90±11 μg/L in the parental strain but did not significantly change the concentrations of 2-aminoadipate, 2-oxopimelate, adipate semi-aldehyde, or adipate.

As the engineered GB20 strain had a slightly higher titer of adipic acid (125±69 μg/L) than the engineered W29 strain (58±46 μg/L), we carried out further metabolic engineering work on the GB20 strain.



	W29 background		GB20 background				
	ST6512	**ST8070**	**ST7806**	**ST8071**	**ST8071_mut**	**ST8067**	**ST8485**
↑*YALI0F31075* homocitrate synthase		+		+		+	+
↑*YALI0F31075 Q377 –> R377*					+		
↑*YALI0E02728g* homoaconitate hydratase		+		+	+	+	+
RK21_02870 aldehyde dehydrogenase						+	+
ChnE aldehyde dehydrogenase						+	+
↑*YALI0F26323* 2-oxoadipate transporter							+
↑*YALI0D02629 Tri-carboxylic acids transporter*							+

Figure 5.3: Production of adipic acid and lysine pathway intermediates by *Y. lipolytica* in mineral medium. ST6512 is a W29 strain with integrated *cas9* and deletion of *ku70Δ*. ST7806 is GB20 engineered in the same way as ST6512. ST8070 and ST8071 strains were made from correspondingly ST6512 and ST7806 by overexpressing homocitrate synthase YALI0F31075p (E.C.2.3.3.14) and homoaconitate hydratase YALI0E02728p (E.C.4.2.1.36). ST8071_mut is analogous to strain ST8071, but it oveexpresses homocitrate synthase with Q377R mutation. ST8067 was made from ST8071 by expressing two heterologous codon-optimized semi-aldehyde dehydrogenases (E.C.1.2.1.63) from *Acinetobacter* (ChnE), and *Pseudomonas* (RK21_02870). ST8485 further overexpresses mitochondrial transporters

for di- and tri-carboxylic acids YALI0D02629p and YALI0F26323p. Data are presented as mean with SD, 4≤N≤14. * stands for p<0.05 in the t-test.

The regulation of lysine biosynthesis in *Y. lipolytica* is not well studied, but in *S. cerevisiae* the lysine pathway is feedback regulated by lysine. The expression of six enzymes of the pathway is 2-4-fold lower in the presence of lysine[37], which represses transcriptional activators[38, p. 14] and activates suppressors[39, p. 80]. In the current study, the homocitrate synthase *YALI0F31075g* (E.C.2.3.3.14), and homoaconitate hydratase *YALI0E02728g* (E.C.4.2.1.36) were overexpressed under the control of strong constitutive promoters, pEXP, and pGPD respectively, and therefore were assumed to not be a subject to a transcriptional/translational regulation by lysine and its precursors.

However, the native homolog of homocitrate synthase in *S. cerevisiae* has a lysine-sensing mechanism, which can be blocked by Q366R mutation in *LYS21*[39]. After aligning Lys21p of *S. cerevisiae* to the *Y. lipolytica* homolog (YALI0F31075p), the analogous mutation Q377R in *Y. lipolytica*'s gene was introduced in an engineered GB20 strain resulting in a strain, which carried overexpressed mutated *YALI0F31075g* (homocitrate synthase) and *YALI0E02728g* (homoaconitate hydratase). However, no statistically significant effects on 2-aminoadipate, 2-oxopimelate, adipate semi-aldehyde, and adipate production were observed, hence the native form of homocitrate synthase was used in the further strain design.

As the next step, two heterologous NAD-dependent adipate-semialdehyde dehydrogenases from *Pseudomonas plecoglossicida (PPJ- RK21_02870)*[40] and *Acinetobacter sp (ChnE)*[41] were codon-optimized and introduced into the engineered GB20 strain, generating a strain which carries overexpressed *YALI0F31075g* (homocitrate synthase) and *YALI0E02728g* (homoaconitate hydratase) and the two heterologous adipate-semialdehyde dehydrogenases. No statistically significant changes in concentrations of adipic acid or its precursors were observed.

Nevertheless, the mean adipate semi-aldehyde, 2-oxopimelate, and 2-oxoadipate peak areas and concentrations were slightly higher in this strain ($3.9 \cdot 10^6$ AU vs $0.13 \cdot 10^6$ AU; $5.6 \cdot 10^6$ AU vs $4.2 \cdot 10^6$ AU; and 571 µg/L vs 413 µg/L respectively). Therefore, the strain carrying both overexpressed native enzymes and heterologous aldehyde dehydrogenases was used for further study.

To increase the exchange of compounds between mitochondria and cytosol, we selected two native mitochondrial transporters: citrate transporter (*YALI0F26323g*), the homolog of which has been recently reported to be involved in lysine biosynthesis in yeast[42], and transporter of 2-oxoglutarate/2-oxoadipate (*YALI0D02629g*), the homolog of which has been reported to be active on 2-oxopimelate as well[43]. The transporters were overexpressed generating a strain which carried overexpressed homocitrate synthase, homoaconitate hydratase, two heterologous adipate-semialdehyde dehydrogenases, and the two

overexpressed transporters. The final strain accumulated 177±67 μg/L of adipic acid, 1.6±0.3 mg/L of 2-oxoadipate. It had a significantly higher amount of adipate semialdehyde than the parental strain, but no longer secreted any 2-oxopimelate (Figure 5.3).

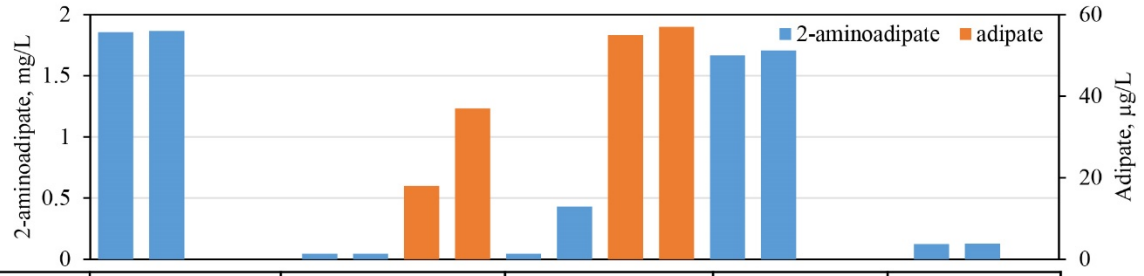
5.2.3.3 Production of adipic acid via the lysine pathway in *Saccharomyces cerevisiae*

Overexpressing the native mitochondrial transporters for di- and tricarboxylic acids in *Y. lipolytica* boosted the accumulation of adipic acid and the key metabolites of the adipic acid pathway in the broth indicating that the availability of the substrates for reactions in the right compartment was important. Therefore, the idea to relocate the enzymatic machinery of the pathway into one compartment (mitochondria or cytosol) was tested in the model organism *S. cerevisiae*, where the signaling peptides for relocating enzymes are known and well-studied[44].

In *S. cerevisiae*, the lysine biosynthesis employs two paralogs of homocitrate synthase Lys20p and Lys21p, homoisocitrate dehydrogenase Lys12p, and homoaconitase Lys4p. The latter is annotated as an enzyme responsible for both reactions: dehydrogenation and hydratation. However, Aco2p was shown to be active in the lysine pathway and controlling the dehydration step[45], and therefore was also included in the list of overexpressed enzymes. Thus, three strains were derived from CEN.PK113-7D parent strain: i) with natively localized overexpressed Lys21p, Aco2p, Lys4p, , and Lys12p; ii) with all 4 enzymes overexpressed in the cytoplasm, where Aco2p, Lys4p, and Lys12p are truncated, lacking their N-terminal mitochondrial signal; and iii) with all 4 enzymes overexpressed in the mitochondria, where Lys21p is fused with a mitochondrial signal from Hsp60p.

No adipic acid, adipate semi-aldehyde, or 2-oxopimelate were detected in any of the strains grown on glucose. We then cultivated the same strains in medium supplemented with 0.2% of 2-oxoglutarate, a precursor of lysine biosynthesis pathway. In this medium, the overexpression of natively localized enzymes Lys21p, Aco2p, Lys4p, and Lys12p decreased the 2-aminoadipate accumulation down to 0.05 mg/L compared to 1.8 mg/L in the control strain CEN.PK113-7D and led to the detection of 25 μg/L of adipic acid, illustrating that the adipic pathway is active not only in *Y. lipolytica*, but in *S. cerevisiae* as well (Figure 5.4). Introducing the feedback resistance point mutation Q366R in overexpressed Lys21p in the strain with natively localized overexpressed enzymes doubled the adipic acid titer.

Targeting overexpressed Aco2p, Lys4p, and Lys12p into the cytosol by truncation of their N-terminal mitochondrial signals did not change the production of 2-aminoadipate or adipic acid, whilst overexpressing all 4 enzymes in mitochondria reduced the 2-aminoadipate accumulation in the broth and no adipic acid was detected. 2-oxoadipate remained at a constant level of 56±0.3 μg/L for all strains.



	Parent	ST8174	ST8174_mut	ST8172	ST8176
Lys21p	Not overexpressed	Overexpressed native form	Oversexpressed with Q366R mutation	Overexpressed native form	Overexpressed fused with Hsp60p
Aco2p	Not overexpressed	Overexpressed native form	Overexpressed native form	Overexpressed trancated	Overexpressed native form
Lys4p	Not overexpressed	Overexpressed native form	Overexpressed native form	Overexpressed trancated	Overexpressed native form
Lys12p	Not overexpressed	Overexpressed native form	Overexpressed native form	Overexpressed trancated	Overexpressed native form

Figure 5.4: Expression of lysine biosynthetic pathway enzymes in different compartments in *S. cerevisiae.* Parent strain is CEN.PK113-7D; ST8174 is a strain carrying overexpressed natively localized Lys21p, Aco2p, Lys4p, and Lys12p. ST8174_mut is its derivative carrying a point mutation (Q366R) in Lys21p; ST8172 carries all four enzymes overexpressed in the cytoplasm; and ST8176 carries all 4 enzymes overexpressed in mitochondria. Data are presented for individual measurements.

The experimental data for *S. cerevisiae* suggests that the native localization of overexpressed enzymes involved in lysine biosynthesis is optimal, therefore, we did not attempt engineering of enzyme localization in *Y. lipolytica*.

5.2.3.4 Production of adipic acid on food waste hydrolyzate by *Y. lipolytica*

The engineered *Y. lipolytica* strain was tested in food waste hydrolysates (Table 5.3). Food waste hydrolysate on average contained 48.8±6.5 g/L of glucose and 7.2±3.2 g/L of xylose as the two major carbon sources, supported by undefined combinations of nutrients, vitamins, and minerals. After 72 hours, glucose was consumed in all samples and adipic acid was produced in amounts from 8.6 mg/L and up to 30.4 mg/L depending on the batch of food waste hydrolysate, with average being 16±5 mg/L compared to 177±67 µg/L in mineral medium with 20 g/L glucose. This corresponds to 90-fold increase of adipic acid titer in the food waste hydrolysate compared to mineral medium.

Table 5.3: Adipic acid from food waste hydrolysate produced by engineered *Y. lipolytica* strain ST8485. WH1, WH2, WH3.1, WH4.2, and WH4.3 are samples of waste hydrolysate prepared according to the scheme described in material and methods. Data are presented as an average of duplicates.

Concentrations	WH1	WH2	WH3.1	WH3.3	WH4.2	WH4.3	Average
Glucose*, g/L	57.7	41.8	52.4	48.4	51.92	40.87	48.85
Xylose*, g/L	9.5	11.3	8.2	7.6	3.4	3.3	7.22
Arabinose*, g/L	0.9	0.5	0.6	0.6	0.1	0.1	0.47
Galactose*, g/L	0	0.5	0.3	0.4	0	0.1	0.22
Maltose*, g/L	5.4	5.7	10.8	10.4	0.4	0.3	5.50
Lactic acid*, g/L	15.9	12.7	5.8	7.3	0.4	1	7.18
Adipic acid**, mg/L	26.71	12.75	13.60	12.43	15.12	15.83	16.07

* Initial concentrations

** End time-point concentration

5.2.4 Discussion

The current work provides evidence that adipic acid can be produced natively by yeasts via the lysine biosynthesis pathway due to the promiscuity of enzymes. The maximum theoretical yield of this pathway is 0.567 g adipate per g glucose. This number is lower than the yield of *de novo* biosynthetic pathway as in *Thermobifida fusca* (0.74 g/g[46]), but the bacterial pathway relies on enzymes from the reverse adipate degradation pathway. Those enzymes are Fe/S cluster dependent, and therefore difficult to transfer into eukaryotic cells. To the best of our knowledge, this bacterial pathway has not yet been successfully expressed in yeast.

The ability of *Y. lipolytica* to produce adipic acid was suggested after the conventional literature and databases search and supported by BLASTp and BrigIT analysis. The latter is a computational tool for finding the most promising candidates for orphan reactions, based on substrate and product chemistry, reactive centers of molecules, and enzyme structures[24]. Its outcome is a score that reflects the probability that a suggested enzyme can perform the orphan reaction. The higher the score (max 1), the more promising the suggested enzyme.

In the current analysis, it was shown that the elongation from 2-oxoadipate to 2-oxopimelate (which was not described previously in yeast) is similar to elongation from 2-oxoglutarate to 2-oxoadipate (which is a part of lysine biosynthesis) and probably can be performed by native enzymes with the BrigIT score between 0.72 to 0.77.

Experimentally, it was proven by comparing two *Y. lipolytica* strains – one with functioning, the other with malfunctioning lysine biosynthesis. Unlike the lysine-auxotrophic strain, the prototrophic strain could produce a trace amount of adipic acid from glucose as the sole carbon source as detected by LC-MS.

In contrast to *Y. lipolytica,* a prototrophic *S. cerevisiae* could not produce adipic acid without supplementing media with 2-oxoglutarate and overexpressing the lysine biosynthesis enzymes Lys21p, Aco2p, Lys4p, and Lys12p under the strong constitutive promoters, which are not sensitive to translation/transcriptional feedback regulation by lysine. Further removal of lysine feedback inhibition by introducing a Q366R mutation in Lys21p[39] increased the concentration of adipic acid in the broth, indicating the importance of lysine regulation for adipic acid accumulation. However, the highest concentration of adipic acid in broth we could achieve in *S. cerevisiae* was below 60 µg/L in media supplemented with 2-oxoglutarate as a pathway precursor, compared to 600 µg/L of adipic acid in non-engineered *Y. lipolytica* strain in mineral media with glucose as a sole carbon source. It may be due to a higher production of 2-oxoglutarate in *Y. lipolytica* [31], [47]–[49].

In *Y. lipolytica*, the lysine feedback inhibition is not as well studied as in *S. cerevisiae*. The analogous mutation Q377 to R377 in *Y. lipolytica*'s *YALI0F31075* (LYS21 analog) found by simple alignment did not give any significant effect on the production of adipic acid or its intermediates. Thus, other mechanisms of lysine feedback inhibition removal should be studied in detail in *Y. lipolytica* for further engineering of strains.

Further engineering of *Y. lipolytica* included introducing heterologous adipate semi-aldehyde dehydrogenases and overexpressing the mitochondrial transporters for di- and tricarboxylic acids. We achieved 177±67 µg/L of adipic acid in mineral medium with glucose as the sole carbon source. The same strain produced up to 30 mg/L adipic acid in food waste hydrolysate.

While the presented pathway allows direct production of adipic acid from sugars and is operational in yeasts, it would require extensive further engineering to increase the titer, rate, and yield.

5.2.5 Conclusion

In the current study, we discovered a novel pathway to adipic acid, which is a subject for further optimization, but was shown to be active in both *Y. lipolytica* and *S. cerevisiae*. The highest titer, about 30 mg/L, was achieved on food waste hydrolysate by a strain carrying overexpressed homocitrate synthase, homoaconitate hydratase, citric and keto-dicarboxylic acids mitochondrial transporters together with adipate semi-aldehyde dehydrogenases from *Acinetobacter* and *Pseudomonas sp*.

5.3 A computational workflow for the expansion of noscapine heterologous biosynthetic pathways to natural product derivatives.

This subchapter is the result of a collaboration with the experimental lab of Prof. Christina Smolke at the University of Stanford. The results of this collaboration led to development of a computational workflow to identify potential derivatives of intermediates of a given biosynthetic pathway and subsequently predict enzyme candidates that may carry out the desired transformation(s). We confirmed the performance of workflow by predicting pathway and enzyme candidates capable of producing (S)-tetrahydropalmatine. This subchapter has been recently submitted as a manuscript for publication. This project was led by Dr. Jasmin Hafner and Dr. James Payne. Dr. Jasmin Hafner used BNICE.ch to analyze the metabolic neighborhood of Noscapine biosynthesis pathway. Experimental results have been obtained by Dr. James Payne, and enzyme prediction using BridgIT tool have been provided by the author of this thesis. Prof. Vassily Hatzimanikatis and Prof. Christina Smolke supervised the project as well as the completion of the manuscript.

Full list of authors in this paper: J. Hafner†, J. Payne†, H. MohammadiPeyhani, V. Hatzimanikatis, and C. Smolke*," A computational workflow for the expansion of heterologous biosynthetic pathways to natural product derivatives" († contributed equally,* corresponding author).*

5.3.1 Introduction

Plants synthesize a remarkable range of complex and valuable molecules, known as plant natural products (PNPs), commonly used as flavors, fragrances, and medicines[50]. However, production of these molecules via extraction from plant biomass is often limited by slow growth, low yield, laborious extraction and purification procedures, and variability due to weather and climate change. Furthermore, while many modern medicines are natural products, a higher fraction are derivatives of natural products[51]. The range of PNP derivatives accessible to researchers is typically limited to those that can be readily produced via chemical synthesis from PNPs extracted from plants, while many more derivatives could potentially be made via regioselective enzymatic modification of PNPs and their intermediates. Microbial production of PNPs can potentially address these concerns, and additionally facilitates production of novel PNP derivatives by leveraging the genetic tractability of well-established microbial hosts to alter the heterologous biosynthetic pathway.

Since the landmark production of the antimalarial drug precursor artemisinic acid in *Saccharomyces cerevisiae* in 2006[52], there has been an increase in the size and complexity of pathways reconstructed in heterologous hosts.[53] This progress is highlighted by the recent *de novo* biosynthesis in *S. cerevisiae* of noscapine[54], an antitussive benzylisoquinoline alkaloid and potential chemotherapeutic[55]–[57] from *Papaver somniferum* separated by 16 enzymatic steps from tyrosine. In that study, halogenated derivatives of tyrosine were fed to the engineered yeast strains to produce halogenated derivatives of noscapine intermediates. However, the non-native halogenated substrates were not tolerated as well as the native substrates of the pathway enzymes, and derivatives of only early intermediates in the pathway were

detected. In such cases, an alternative strategy would be required to produce derivatives of more chemically complex downstream pathway intermediates or of noscapine itself.

An alternative approach to produce derivatives of PNPs and their intermediates is to integrate additional enzymes into microorganisms expressing heterologous PNP biosynthetic pathways. Enzymes that are able to accept and functionalize intermediates or products along a PNP pathway would thus produce novel products *in vivo* from the natural precursors. However, producing new-to-nature compounds necessarily entails the use of enzymes outside their natural functions (promiscuous activity). Developing computational tools able to learn from the wealth of enzymatic knowledge and predict new catalytic promiscuity will be of great value.

Computational methods have been employed to guide the discovery of enzymatic functions and the design of biosynthetic pathways for the production of molecules with interesting pharmaceutical or industrial properties[58]. These methods generate hypothetical pathways to compounds of interest by assuming that enzymes that perform similar, but not identical, reactions to those desired might be promiscuous or sufficiently evolvable to perform the desired reaction after engineering and/or optimization. The concept of substrate promiscuity is translated into *generalized enzymatic reaction rules* that mathematically describe the reactive site recognized by an enzyme as well as the molecular rearrangement performed during the biotransformation. Popular cheminformatic tools[58]–[60] for predictive biochemistry include BNICE.ch (Biochemical Network Integrated Computational Explorer)[61], enviPath[62], GEM-Path[63], NovoPathFinder[64], NovoStoic[65], ReactPRED[66], RetroPath2.0[67], and Transform-MinER[68]. These tools have typically been used in retrobiosynthesis studies, where the aim is to determine potential bioproduction pathways by biochemically walking back from a target compound to the native metabolism of a chassis organism[69]–[71] via predicted enzymatic reaction steps. The prediction of novel reactions is subsequently followed by the search for suitable enzymes that can catalyze the predicted step. Enzyme prediction tools such as BridgIT[4], EC-BLAST[5], E-zyme[6] and Selenzyme[7] determine the structural similarity of a novel reaction to all well-characterized reactions in biochemical databases, and propose a list of enzyme candidates ranked by their likelihood to catalyze the desired transformation.

Here, we develop a computational workflow to identify potential derivatives of intermediates of a given biosynthetic pathway and subsequently predict enzyme candidates that may carry out the desired transformation(s) (Figure 5.5). In contrast to previously reported retrobiosynthesis studies, in which a predicted pathway to a given target is generated, our workflow begins with a set of starting compounds (i.e., the intermediates of a heterologous biosynthetic pathway) and determines a suite of novel target compounds and associated pathways that can be generated. The method expands the chemical space around a pathway of interest using BNICE.ch to create a map of all compounds accessible with known biochemical reactions and then identifies enzymes capable of carrying out the desired transformations on the prioritized set of compounds using the enzyme prediction tool BridgIT. We applied this workflow to the reconstructed

noscapine biosynthetic pathway in yeast. We narrowed our search to enzyme candidates capable of producing *(S)*-tetrahydropalmatine, a PNP found in plants of the genus *Corydalis* that has been shown to possess analgesic and anxiolytic effects and has shown promise as a potential treatment for opiate addiction[72]–[74]. After experimental evaluation of seven of the top enzyme candidates in yeast strains engineered to produce the noscapine biosynthetic intermediate *(S)*-tetrahydrocolumbamine *de novo*, two enzymes were identified that enabled production of *(S)*-tetrahydropalmatine. To our knowledge, our work describes the first use of a computational workflow to expand a heterologous biosynthetic pathway to produce additional compounds. As the number of reconstructed heterologous pathways for PNPs continues to increase, we anticipate that the described workflow can be used to produce many chemically complex compounds spanning diverse therapeutic activities.

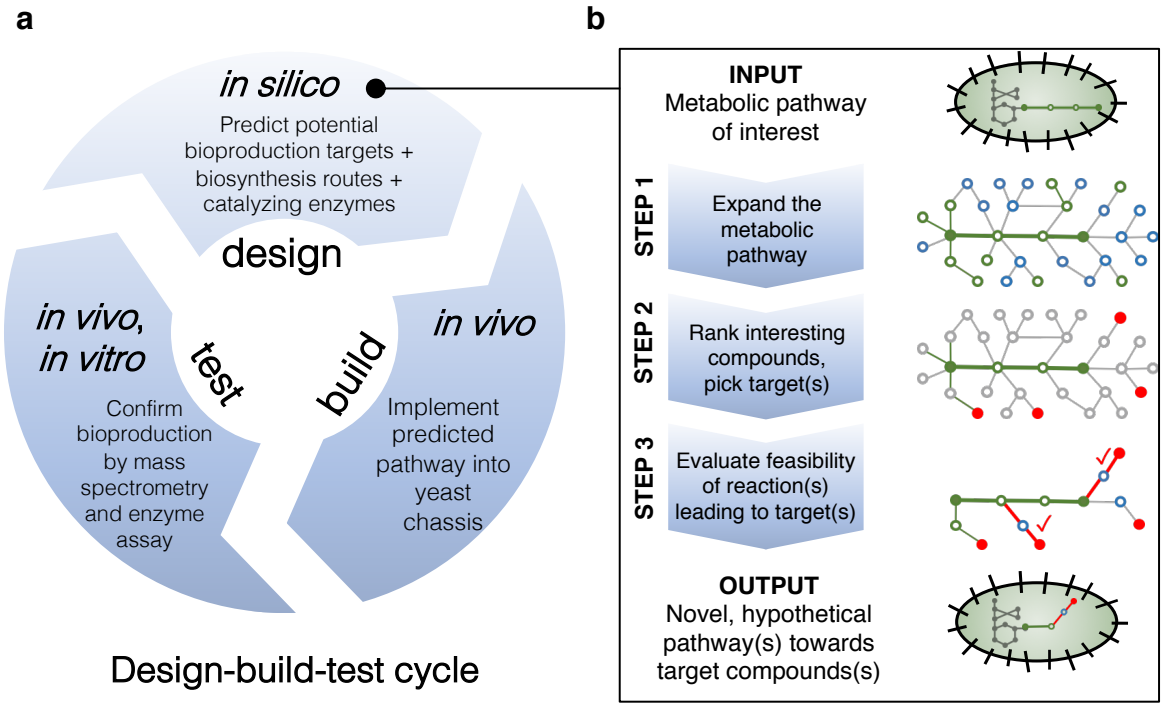


Figure 5.5: Overall workflow integrating computational prediction of target compounds, pathways, and enzymes with experimental validation. a) Applied design-build-test cycle. b) Computational workflow. Circles represent compounds, edges represent biotransformations. Green is used to designate known biological reactions and compounds, blue circles are compounds from the chemical space without specific biological annotation, and red circles show compounds selected for their popularity in scientific literature and in the patent landscape.

5.3.2 Materials and Methods

5.3.2.1 Computational exploration of the biochemistry surrounding the noscapine pathway

The computational workflow consists of three steps: (i) expansion of a biochemical reaction network around the original pathway, (ii) popularity assessment of compounds via annotation and ranking, and (iii) feasibility assessment via reaction annotation, pathway assembly, and pathway evaluation. The output of the computational analysis was directly used for the design of engineered yeast strains.

5.3.2.2 Expansion of a biochemical network

A hypothetical biochemical network using BNICE.ch[75] was expanded around the input pathway, consisting of 17 metabolites connected by 17 reactions and catalyzed by a total of 11 generalized reaction rules, using a collection of 442 bidirectional generalized enzymatic reaction rules. In a first iteration, the integrated network generation algorithm applies the reaction rules on the input molecular structures (MDL molfiles), which generates all biochemically possible reactions according to the reaction mechanisms represented in BNICE.ch. The products of these reactions are stored, and used as input compounds for the next iterations of reaction generation. This iterative process generates hypothetical biochemical networks around any given set of input molecules.

BNICE.ch distinguishes between known and novel compounds by looking up the generated molecular structures in different databases: if the compound is part of any biological, bioactive, or chemical database it is considered as known and annotated with the corresponding database identifiers. The following databases are used: the Kyoto Encyclopedia of Genes and Genomes (KEGG)[76], SEED[77], HMDB[78], MetaCyc[79], Brenda[80], MetaNetX[81], Rhea[82], BiGG[83], PMN[84], KNApSAcK[85] for biological compounds, ChEBI[86] and ChEMBL[87] for bioactive compounds, and PubChem[88] for chemical compounds. In this workflow, only known molecular structures are allowed in the network generation. Reactions are classified as known if they are part of the KEGG reaction database or the noscapine pathway, and as novel if they are not.

5.3.2.3 Compound annotation and ranking

We assessed the “popularity” of the generated compounds in the second step of the workflow by determining how many times each compound appears in scientific publications, and how many patents are associated with the molecule. The number of publications was derived from PubChem and PubMed, while the number of patent annotations was extracted from PubChem. We used the PUG-REST service to retrieve information on compounds from the PubChem website (https://pubchem.ncbi.nlm.nih.gov/)[89] on the number of associated patents and citations. We also used the Entrez Programming Utilities (E-utilities) API service to search the PubMed database for citations by compound name[90]. We only kept compounds with at least one annotation as potential targets for biosynthesis.

5.3.2.4 Reaction annotation and pathway ranking

To determine if the potential targets for biosynthesis have valid bioproduction pathways, we listed all possible pathways connecting any noscapine pathway intermediate to the potential target within a maximum of four reaction steps. A path search algorithm was employed to extract linear pathways from the network of reactant-product pairs. Reaction directionalities producing molecular oxygen and reverse decarboxylations were excluded from the pathway search because of their high energy demand. Also, demethylation reactions producing S-adenosylmethionine from S-adenosylhomocysteine were not allowed (other demethylation transformations were allowed).

To find enzymes for the predicted reactions in each pathway, we used the enzyme prediction tool BridgIT[4]. BridgIT calculates a similarity score between the novel reaction and reactions from a reference database of known, enzyme-annotated reactions (KEGG reaction database, downloaded in February 2018) by comparing the molecular fingerprints on and around the reactive sites of the participating reactants. The similarity between fingerprints is expressed as a score ranging from zero (no similarity) to one (the two reactions are identical up to seven atoms around the reactive site). A BridgIT score above 0.3 is considered as significant. For each reaction in the pathways, we performed BridgIT and we collected all the reactions from the reference database that had a score of 0.3 or higher. From the top score of each reaction in the pathway, we calculated the average to provide an overall metric for the enzymatic feasibility of the pathway. The pathways are available online including the top five enzymes predicted by BridgIT and associated similarity scores.

5.3.2.5 Yeast strain construction

Strains used in this work are listed in Supplementary Table 8.10. All strains used are derived from the previously reported strain CSY1171[91]. Strains were grown non-selectively in yeast-peptone media supplemented with 2% w/v dextrose (YPD media), yeast nitrogen base (YNB) defined media (Becton, Dickinson and Company, BD) supplemented with synthetic complete amino acid mixture (YNB-SC; Clontech) and 2% w/v dextrose, or on agar plates made using the aforementioned media. Strains transformed with plasmids bearing the URA3 auxotrophic selection marker were grown selectively in YNB media supplemented with 2% w/v dextrose and uracil (YNB-Ura; Clontech) or on YNB-Ura agar plates.

Yeast genomic modifications were performed using the CRISPRm method[92]. Oligonucleotides used in this work (Appendix, Table 8.11) were synthesized by the Stanford Protein and Nucleic Acid Facility (Stanford, CA). Biosynthetic genes used in this study (Appendix, Table 8.12) were codon-optimized using GeneArt Gene Optimizer software (Thermo Fisher Scientific) either for expression in *S. cerevisiae* or *E. coli* (Appendix, Table 8.15) and then synthesized as either gBlock DNA fragments (Integrated DNA Technologies, IDT) or gene fragments (Twist Bioscience). All biosynthetic genes were synthesized with overhangs on both the 5’ end (5’ – TCGACGGATTCTAGAACTAGTGGATCCTATACA – gene – 3’) and 3’ end (5’ – gene – TAGCCATAAGAATTCAGACACTCGAGAACTCA – 3’) for ease of cloning. CRISPRm plasmids expressing

Streptococcus pyogenes Cas9 (SpCas9) and a single guide RNA (sgRNA) targeting a locus of interest in the yeast genome were constructed by assembly PCR and Gibson assembly of DNA fragments encoding SpCas9 (pCS3410), tRNA promoter and HDV ribozyme (pCS3411), a 20-nt guide RNA sequence (synthesized by the Stanford Protein and Nucleic Acid Facility), and tracrRNA and terminator (pCS3414)[93]. For gene insertions, integration fragments containing the gene(s) of interest flanked by a promoter and terminator were constructed by PCR amplification such that they possessed 40 bp overhangs on either end with homology to the yeast genome surrounding the site targeted by the guide RNA sequence. Approximately 300 ng of each integration fragment was co-transformed with 300 ng of the CRISPRm plasmid expressing the sgRNA targeting the desired genomic site. Positive integrants were identified by yeast colony PCR, DNA sequencing (Quintara Biosciences; South San Francisco, CA), and/or functional screening by LC-MS.

5.3.2.6 Plasmid construction

Plasmids used in this study (Appendix, Table 8.13) were constructed through Gibson assembly. Gibson assembly was performed by amplifying both the gene of interest and the destination plasmid (pCS952[94] or pET28) with 40 bp homologous overhangs. PCR amplifications were performed using Q5 DNA polymerase (NEB) and linear DNA fragments were purified using the DNA Clean and Concentrator-5 kit (Zymo Research). Assembled plasmids were propagated in chemically competent *E. coli* (TOP10; Thermo Fisher Scientific) using heat-shock transformation and selection on Luria-Bertani (LB)-agar plates with carbenicillin (100 µg/mL; for pCS952 derived plasmids) or kanamycin (50 µg/mL; for pET28 derived plasmids). Plasmid DNA was isolated by alkaline lysis from overnight *E. coli* cultures grown at 37 °C and 250 rpm in selective LB media using Econospin columns (Epoch Life Science) according to the manufacturer's protocol.

5.3.2.7 Yeast transformations

Yeast strains were chemically transformed using the Frozen-EZ Yeast Transformation II Kit (Zymo Research). Individual colonies were inoculated into YPD media and grown overnight at 30 °C and 250 rpm. Saturated cultures were back-diluted into three new cultures at 1:5, 1:10, and 1:20 dilutions in YPD media and grown for an additional 5–7 hours to reach exponential phase. For each transformation, 1 mL aliquots from each back-diluted culture were pelleted by centrifugation at 500 g for 4 minutes (successively pelleting aliquots from each different dilution into a single pellet in a 1.5 mL microcentrifuge tube) and then washed twice by resuspending the pellet in 1 mL of 50 mM Tris-HCl buffer, pH 8.5. Washed pellets were resuspended in 50 µL of EZ2 solution per transformation and mixed with 100–600 ng of total DNA and 500 µL of the EZ3 solution. The yeast suspensions were incubated at 30 °C with gentle inversion for one hour. For plasmid transformations, the transformed yeast were directly plated onto YNB-Ura agar plates. For Cas9-mediated gene integrations, the yeast suspensions in the EZ3 solution were first mixed with 1 mL YPD media, pelleted by centrifugation at 500 g for 4 minutes, and then resuspended in 250 µL of fresh YPD media. The suspensions were incubated at 30 °C with gentle inversion for an additional 90 minutes to allow production of G418

resistance proteins and then spread onto YPD plates containing 400 mg/L G418 sulfate. For all transformations, plates were incubated at 30 °C for 72 hours before being used to inoculate cultures for metabolite assays.

5.3.2.8 Growth conditions for metabolite assays

Metabolite production tests were performed in YNB-SC or YNB-Ura media with at least three replicates. Yeast colonies were inoculated into 300 µL of media and grown in 2 mL deep-well 96-well plates covered with AeraSeal gas-permeable film (Excel Scientific). Cultures were then grown for 72-120 hours (exact duration is specified in each figure) at 30 °C, 460 rpm, and 80% relative humidity in a Lab-Therm LX-T shaker (Adolf Kuhner).

5.3.2.9 Analysis of metabolite production

Cultures were pelleted by centrifugation at 3500 g for 5 minutes at 4 °C and 100 µL aliquots of the supernatant were removed for direct analysis. Metabolite production was analyzed by LC-MS/MS using an Agilent 1260 Infinity Binary HPLC and an Agilent 6420 Triple Quadrupole mass spectrometer. Chromatography was performed using a Zorbax EclipsePlus C18 column (2.1 × 50 mm, 1.8 µm; Agilent Technologies) with water with 0.1% v/v formic acid as solvent A and acetonitrile with 0.1% v/v formic acid as solvent B. The column was operated with a constant flow rate of 0.4 mL/minute at 40 °C and a sample injection volume of 5 µL. Compound separation was performed using the following gradient: 0.00–0.10 minutes, 10% B; 0.10–5.00 minutes, 10-40% B; 5.00–5.50 minutes, 40% B; 5.50–6.00 minutes, 40–98% B; 6.00–10.00 minutes, 98% B; 10.00–10.01 minutes, 98-10% B; 10.01–13.00 minutes, equilibration with 10% B. The LC eluent was directed to the MS from 1–10 minutes operating with electrospray ionization (ESI) in positive mode, source gas temperature 350 °C, gas flow rate 11 L/minute and nebulizer pressure 40 psi. Metabolites were quantified by integrated peak area in MassHunter Workstation software (Agilent) based on the multiple reaction monitoring (MRM) parameters in Appenddix Table 8.14. Integrated peak areas were converted to titers by comparison to standard curves prepared using a commercial standard of *(S)*-tetrahydropalmatine (Toronto Research Chemicals). Primary MRM transitions for *(S)*-tetrahydropalmatine were identified by analysis of a 0.1 mM standard in methanol using the MassHunter Optimizer software package (Agilent); all other MRM transitions used were previously reported[94].

5.3.2.10 Enzyme expression and purification

Plasmids containing the gene of interest in a pET28 expression vector (see Appenddix, Table 8.13 for a full list of plasmids used in this study) were used to transform *E. coli* BL21(DE3) (Invitrogen) competent cells containing the pGro7 chaperone expression plasmid (Takara) via heat shock. Briefly, 1 ng of plasmid DNA was added to a 50 µL aliquot of competent cells, the tube was chilled on ice for 15 minutes, placed in a 42 °C

water bath for 35 seconds, then returned to ice for 2 minutes. Seven hundred fifty µL of SOC media were then added and the tube was rotated at 37 °C for 45 minutes before being plated on an LB agar plate containing 50 µg/mL kanamycin and 20 µg/mL chloramphenicol. A single colony was then picked and used to inoculate a primary culture of 5 mL of LB media containing 50 µg/mL kanamycin and 20 µg/mL chloramphenicol which was then grown for 24 hours. Five hundred µL of this primary culture were then used to inoculate a secondary or expression culture of 50 mL of TB medium containing 50 µg/mL kanamycin and 20 µg/mL chloramphenicol (for all proteins except PsS9OMT) or 500 mL of LB medium containing 50 µg/mL kanamycin (for PsS9OMT). This expression culture was grown to an OD_{600} of 0.6-1.0 and then induced with IPTG (for O-methyltransferase induction, GoldBio) and L-arabinose (for groES/groEL induction, Fischer Scientific) at final concentrations of 0.1 mM and 2 mg/mL, respectively, for all proteins except PsS9OMT, which was induced with only IPTG to a final concentration of 1 mM. The expression culture was then grown at 30 °C (for all proteins except PsS9OMT) or 16 °C (for PsS9OMT) for 20 hours at 250 rpm, after which, the culture was harvested by centrifugation (10 minutes at 3,500 rpm in a 50 mL Falcon tube) and stored at -20 °C until lysis and purification.

Frozen pellets were then thawed and resuspended in 25 mL of Ni-nitrilotriacetic (Ni-NTA) equilibration buffer (50 mM sodium phosphate, 300 mM NaCl, 10 mM imidazole, pH 7.4) and lysed by sonication while kept on ice (Branson Sonifier 450, 0.5” horn, 50% duty cycle, 4 x 1 minute with 2 minute rests). Lysed cultures were then clarified by centrifugation (45 min at 35,000 g at 4 °C) and the clarified lysate was purified by Ni-NTA affinity chromatography. Briefly, 1 mL of Ni-NTA resin (Fisher Scientific) was equilibrated with at least 5 volumes of Ni-NTA equilibration buffer (described above) and then loaded with the clarified lysate. The loaded resin was then washed with at least 5 volumes of Ni-NTA wash buffer (50 mM sodium phosphate, 300 mM NaCl, 50 mM imidazole, pH 7.4) and then the bound protein was eluted with 5 volumes of Ni-NTA elution buffer (50 mM sodium phosphate, 300 mM NaCl, 250 mM imidazole, pH 7.4). The eluted fractions were then combined and concentrated using an Amicon® 30 kDa cutoff spin filter (EMD Millipore) at 5,000 g at 4 °C. Concentrated protein fractions were then exchanged into storage buffer (50 mM potassium phosphate, 100 mM NaCl, 10% glycerol, pH 7.5), split into separate aliquots, and stored at -20 °C until use.

5.3.2.11 *In vitro* bioconversions

Analytical reactions were carried out at the 50 µL scale in triplicate. To a 1.5 mL Eppendorf tube were added 5 nmol substrate (final concentration of 100 µM; *(S)*-norcoclaurine and *(S)*-scoulerine purchased from Toronto Research Chemicals; norlaudanosoline purchased from Santa Cruz Biotechnology), 1 µmol sodium ascorbate (final concentration of 25 mM), 5 nmol S-adenosylmethionine (SAM, final concentration of 100 µM; purchased from Sigma-Aldrich), and 150 pmol purified methyltransferase enzyme (3 µM final concentration) in 50 mM potassium phosphate, pH 8.0. The reactions were shaken at 600 rpm at 37 °C for 2 hours before being quenched with an equal volume of methanol, spun down at 20,000 g for 10 minutes, and

filtered prior to LC-MS analysis (see “Analysis of metabolite production” section above for details on LC-MS analysis conditions).

5.3.2.12 Metabolite purification

The large scale *in vitro (S)*-scoulerine conversion reaction was carried out on the 20 mg scale at a final reaction volume of 610 mL in a 2 L Erlenmeyer flask. To this flask were added 20 mg (61 µM) of *(S)*-scoulerine (final concentration of 100 µM), 15 mmol sodium ascorbate (final concentration of 25 mM), 61 µM SAM (final concentration of 100 µM), and 73 nmol of purified TfS9OMT DS M111A[91] (final concentration of 0.12 µM) in 50 mM potassium phosphate, pH 8.0. The reaction was incubated at 37 °C at 250 rpm. The reaction was ultimately run for 15 hours, but was monitored to ensure conversion had stopped by analytical LC-MS. To do so, 50 µL aliquots were pulled periodically, quenched with an equal volume of MeOH, spun down at 20,000 g for 10 minutes, and filtered prior to LC-MS analysis (see “Analysis of metabolite production” section above for details on LC-MS analysis conditions).

Once the reaction was complete, 30 g of Amberlite XAD4 resin were added and the flask was shaken overnight at 30 °C at 250 rpm. The Amberlite XAD4 resin was transferred to 50 mL Falcon tubes, the supernatant was decanted off, and 50 mL total MeOH were then added to the 4 tubes containing resin. The resin in MeOH was then vortexed for 10 minutes, after which it had turned yellow. The MeOH was then pipetted into a 500 mL round-bottomed flask and was concentrated by rotary evaporation to ~2 mL, which was then pipetted into 4 tared 1.5 mL Eppendorf tubes and concentrated to dryness overnight on a speedvac. Approximately 400 mg of crude material were obtained from this process, which were then resuspended in H_2O to a final concentration of 100 mg crude material/mL. This material was then purified by preparative LC (Agilent 1200 Series LC) with a Varian Pursuit XRs C18 250 x 10 mm column, 5 µm particle size (solvent A = H_2O with 0.1% FA, solvent B = ACN with 0.1% FA). The following LC method was used: 0-4.0 minutes, 20% B, 2.0 mL/minute; 4.0-12.0 minutes, 20-100% B, 2.0 mL/minute; 12-20 minutes, 100% B, 2.0 mL/minute; 4 minute postrun. Fractions were analyzed by LC-MS to determine which contained the desired products (see “Analysis of metabolite production” section above for details on LC-MS analysis conditions). Fractions containing the desired product were concentrated and re-purified by preparative LC until the desired purity was obtained.

5.3.3 Results

5.3.3.1 Computational expansion of the noscapine pathway reveals thousands of potential target molecules

Each biosynthetic pathway presents an opportunity to produce numerous derivative compounds by chemically modifying functional groups of the pathway product and its intermediates. Computational reaction prediction tools, such as BNICE.ch, allow rapid exploration of the hypothetical chemical space of potential pathway derivatives. Their generalized enzymatic reaction rules mimic known enzymatic activities

in silico by recognizing and transforming a specific functional group on a substrate to generate a product. Iterative application of these rules to biosynthetic pathway intermediates creates a reaction network to hypothetical derivatives of all pathway intermediates, offering new targets for bioproduction.

We applied this computational expansion process on the noscapine pathway, which starts from *(S)*-norcoclaurine and involves 17 metabolites connected by 17 reactions (Figure 5.6). BNICE.ch expanded the network around the 17 metabolites for four generations, generating both known and novel reactions to produce compounds known to any biological[76]–[79], [81], [82], [84], [85], [95], [96], bioactive[86],[87], or chemical[88] database. This expansion yielded a network spanning 4,838 compounds and 17,597 reactions (Tables 8.16, 8.17). As our analysis focused on BIAs, we required the substrate and product to contain the minimal elemental composition of the 1-benzylisoquinoline scaffold (i.e., at least 16 carbon atoms, 13 hydrogen atoms, and 1 nitrogen atom). The resultant trimmed BIA network spanned 1,518 compounds, of which 99 were classified as biological or bioactive, and the remaining 1,419 as chemical compounds (appendix, Table 8.18). The compounds in the network were connected by 7,527 reactions, of which 49 were known to be catalyzed by well-characterized enzymes linked to a genetic sequence from at least one organism in our reference database, the Kyoto Encyclopedia of Genes and Genomes (KEGG)[76].

Our network expansion was non-uniform across the noscapine biosynthetic pathway (Figure 5.6). The upstream portion of the network is highly connected, whereas the downstream portion near noscapine is less populated. This likely results from the downstream intermediates and their derivatives increasing in size and complexity, complicating their experimental detection and structural characterization. Consequently, these compounds are less represented in biological or chemical databases, and therefore are not part of the predicted network despite their increased diversity of functional groups.

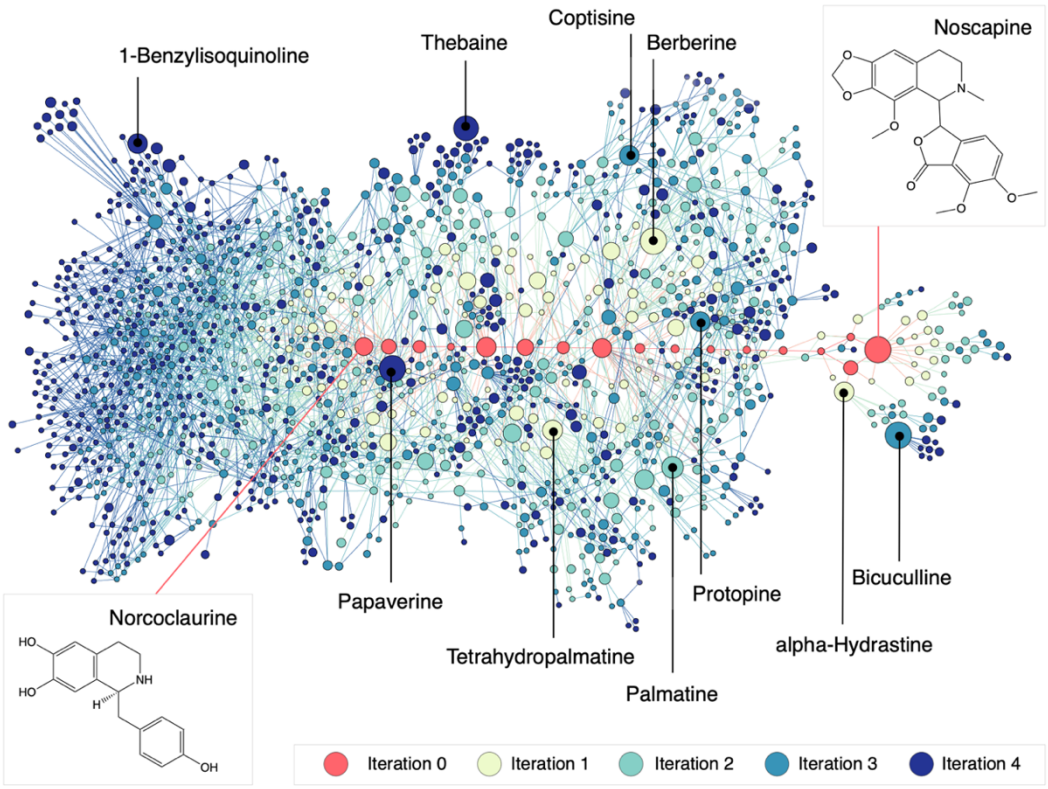


Figure 5.6: Visualization of the expanded biosynthesis network of the noscapine pathway. The nodes and edges drawn in red show the original noscapine pathway. Around the original pathway, the predicted network of compounds (nodes) and reactions (edges) is visualized. The top 10 compounds in terms of popularity (total number of patents plus citations) are named and localized on the map. The color of the nodes shows in which iteration the compound was generated in the network reconstruction process, which is also the number of reaction steps between the original pathway and the compound. The size of the nodes is proportional to the popularity. The molecular structure of the pathway precursor, norcoclaurine, and the final product, noscapine, are shown.

5.3.3.2 A ranking algorithm for candidate molecules highlights well-studied compounds

To guide experimental efforts toward interesting targets for bioproduction, the numerous candidate compounds were ranked and filtered. To focus on compounds with broader interest to biomedical researchers, we ranked the candidates by “popularity”, defined here as the sum of the number of citations and patents reported. We screened the 1,501 potential target compounds (1,518 satisfying the BIA requirement minus the 17 in the noscapine pathway) and found that 204 returned at least one citation, while 467 had at least one associated patent. In total, at least one annotation (citation or patent) was obtained for 545 distinct compounds (appendix, Table 8.19).

Sorting the compounds by popularity, we found that papaverine was ranked highest, with 22,918 annotations, followed by bicuculline and berberine with 16,118 and 12,154 total annotations, respectively. While the citation count reflects scientific interest in a compound, the number of patents indicates its

commercial applications. As an example, the compound bicuculline, which ranked first in citations but fourth in patents, is widely employed in medical research to mimic epilepsy in mammals[97], but has a relative lack of clinical applications.

5.3.3.3 Computational pathway construction identifies tetrahydropalmatine as a high-priority target

While the application of a ranking algorithm to the potential compounds generated by BNICE.ch identifies top candidates, it does not prioritize those which can be feasibly produced experimentally. To maximize the probability of successful *in vivo* production of a target molecule, we applied additional filters to determine the best candidates for bioproduction. Four criteria were considered: (i) one or more production pathways toward the target compound are thermodynamically feasible; (ii) enzymes are available which natively perform similar transformations; (iii) the target compound is only one chemical transformation from an intermediate in the original pathway to focus experimental efforts on a single enzymatic step; and (iv) the target molecule is a potential or confirmed pharmaceutical.

We first examined the biological feasibility of the potential pathways to our target compounds. For the top 50 ranked candidates, we enumerated all possible pathways connecting a noscapine pathway intermediate to each target within a maximum of four reaction steps. Reactions with a high standard Gibbs free energy of reaction (i.e., reactions producing molecular oxygen, binding carbon dioxide to the substrate, or demethylating the substrate via *S*-adenosylhomocysteine) were excluded to avoid thermodynamic and catalytic bottlenecks. We identified feasible pathways for 42 of 50 targets, furnishing a total of 1,338 pathways (appendix Table 8.20). Providing validation of our approach, the known biosynthetic pathway for protopine is included in this set. All of the proposed pathways are listed and visualized online (https://lcsb-databases.epfl.ch/pathways/GraphList).

To assess the availability of enzymes to catalyze the proposed reactions, we predicted enzymes for each novel reaction step using BridgIT[4]. BridgIT calculates a reactive-site centric similarity score (BridgIT score) between the novel reaction and a reference database of known, well-characterized reactions (KEGG). The output is a ranked list of candidate enzyme classes and associated similarity scores that indicate the probability that members of the candidate enzyme class will catalyze the novel reaction. As an overall metric for compound feasibility, we used the mean of the top BridgIT scores of each reaction in the pathway (available as part of the pathway visualization online).

We next examined the distance (i.e., number of reaction steps) of the target compounds from the original pathway. We restricted our search to candidates that are only one reaction from an intermediate, resulting in 15 candidates, each produced by a feasible reaction and associated with a ranked list of predicted, putative enzymes (Table 5.4, appendix Table 8.21). The highest ranked candidate was berberine, for which a heterologous biosynthetic pathway has already been established[98]. We therefore selected the second

highest ranked candidate, *(S)*-tetrahydropalmatine, for experimental validation. *(S)*-Tetrahydropalmatine naturally occurs in a number of plants, especially those in the genus *Corydalis* and *Stephania rotunda*, which are traditionally used in Chinese herbal medicine[99]. *(S)*-Tetrahydropalmatine (i.e., levo-tetrahydropalmatine) has been used for its analgesic, anxiolytic, and sedative effects as an alternative to opiates and benzodiazepines, and has shown promise in treating opiate, cocaine, and methamphetamine addiction [74].

Table 5.4**:** List of compounds ordered by descending popularity that are one reaction step away from intermediates in the noscapine pathway.

Popularity Rank	Name	Best BridgIT score	Predicted EC	Number of Citations	Number of Patents	Citations + Patents
1	Berberine	1.00	1.3.3.8	5430	6751	12154
2	Tetrahydropalmatine	1.00	2.1.1.89	530	355	885
3	Columbamine	0.99	1.3.3.8	131	235	366
4	Salutaridine	1.00	1.14.19.67	85	264	349
5	Norlaudanosoline	0.99	1.14.14.102	144	177	321
6	Stepholidine	0.78	1.14.13.31	157	140	297
7	Allocryptopine	0.32	1.14.13.239	111	159	270
8	Laudanidine	1.00	2.1.1.291	23	112	135
9	Codamine	0.79	2.1.1.121	13	61	74
10	Norreticuline	0.09	1.5.3.10	33	40	73
11	Corytuberine	0.56	1.14.19.67	18	39	57
12	Lambertine	0.45	1.3.1.29	30	23	53
13	Armepavine	1.00	2.1.1.291	28	15	43
14	1,2-Dehydroreticuline	1.00	1.5.1.27	3	40	43
15	Nandinine	1.00	1.14.19.73	1	39	40

5.3.3.4 BridgIT analysis indicates top enzyme candidates for tetrahydropalmatine bioproduction

Once a compound of interest is chosen, enzyme(s) catalyzing the desired transformation must be identified. BridgIT identifies known enzymes whose native reactions most closely resemble our desired reaction, and the BridgIT similarity score can be used to rank the candidates by their likelihood to catalyze the desired transformation.

(S)-Tetrahydropalmatine can be produced in one step via methylation of the 2-hydroxyl of the noscapine pathway intermediate *(S)*-tetrahydrocolumbamine with concomitant conversion of S-adenosylmethionine to S-adenosylhomocysteine (Figure 5.7 panel a). Because of the lack of sequence annotation for this reaction in KEGG, we used the BridgIT data described above to identify candidate enzymes. The BridgIT analysis produced a list of enzyme classes ranked by their BridgIT scores, measuring the structural similarity of the

(S)-tetrahydrocolumbamine methylation to the native reactions of those enzymes (Table 5.5). Enzymes without protein sequence annotation were removed.

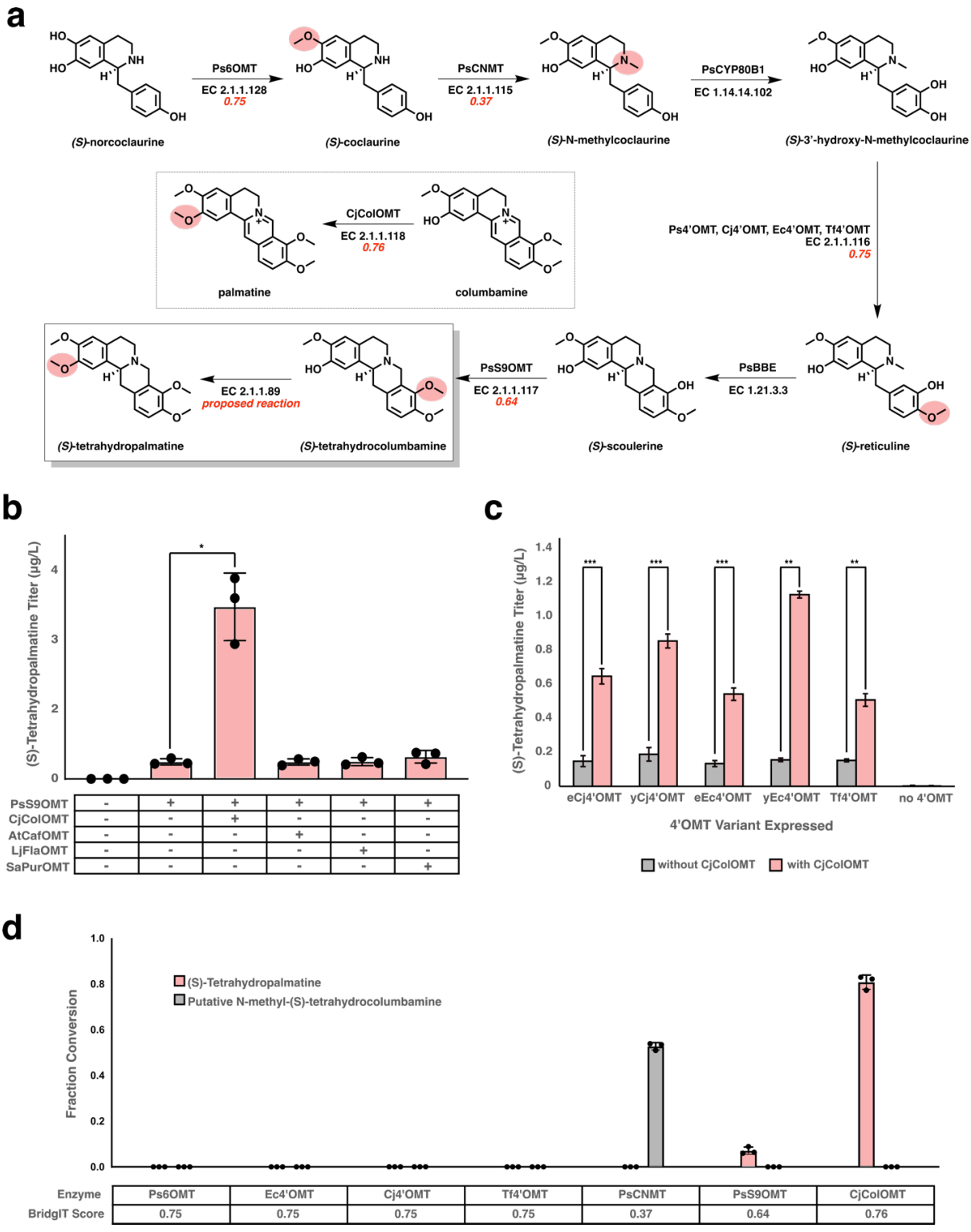


Figure 5.7: In vivo and in vitro activity of predicted enzymes. a) Biosynthetic pathway from (S)-norcoclaurine, the first dedicated intermediate in the pathway, to (S)-tetrahydropalmatine. The specific enzyme(s) used in our strains are indicated above each reaction arrow, while below each arrow is the enzyme class and, for methyltransferases, the BridgIT score (in red) obtained for the likelihood of members of that class to perform our proposed reaction. Our proposed reaction, the methylation of (S)-tetrahydrocolumbamine

to afford (S)-tetrahydropalmatine, is shown in the box at the bottom left. Shown in dotted lines is the native reaction of CjColOMT, the enzyme which was predicted and demonstrated to perform our proposed reaction. The site of methylation of each methyltransferase is highlighted on its product in pink. b) De novo production of (S)-tetrahydropalmatine in yeast strains engineered to express members of the two most downstream O-methyltransferase classes (S9OMT & ColOMT) predicted by BNICE.ch & BridgIT to accept (S)-tetrahydrocolumbamine as a substrate. PsS9OMT is integrated into the yeast genome, while CjColOMT, AtCafOMT, LjFlaOMT, and SaPurOMT were expressed from a high-copy plasmid; the first two strains shown contain an empty version of this plasmid. Strains were cultured in selective media (YNB-Ura) with 2% dextrose, 2 mM L-DOPA, and 10 mM ascorbic acid at 30 °C for 120 hours before LC-MS/MS analysis of the growth media. Asterisks represent Student's two-tailed t-test: *P < 0.05,**P < 0.01, ***P < 0.001. c) In vitro reactions of purified methyltransferases on (S)-tetrahydrocolumbamine to produce (S)-tetrahydropalmatine (shown in pink) or the putative N-methyl-(S)-tetrahydrocolumbamine product (shown in gray). BridgIT score denotes the score obtained by BridgIT for the enzyme class to which each enzyme belongs. d) De novo production of (S)-tetrahydropalmatine in yeast strains engineered to express alternative 4'OMTs. Strains were cultured in selective media (YNB-Ura) with 2% dextrose, 2 mM L-DOPA, and 10 mM ascorbic acid at 30 °C for 72 hours before LC-MS/MS analysis. Asterisks represent Student's two-tailed t-test: *P < 0.05,**P < 0.01, ***P < 0.001.

Table 5.5: Reaction similarities between the predicted tetrahydropalmatine-producing reaction and its top 18 most similar, gene-annotated reactions from the BridgIT reference database.

Rank	BridgIT score	Predicted EC	Native substrate	Type of substrate	Native organism	Enzymes tested	Activity on THCB
1	0.98	2.1.1.291	*(S)*-Reticuline	BIA	*P. somniferum*		Not tested
2	**0.76**	**2.1.1.118**	**Columbamine**	**BIA**	***Coptis japonica***	**CjColOMT**	**Active**
3	0.75	2.1.1.128	*(S)*-Norcoclaurine	BIA	*P. somniferum*	Ps6OMT	Already in pathway, no activity
4	0.75	2.1.1.116	3'-Hydroxy-N-methyl-*(S)*-coclaurine	BIA	*P. somniferum* *C. japonica* *E. californica* *T. flavum*	Ps4'OMT Cj4'OMT Ec4'OMT Tf4'OMT	Already in pathway, no activity
5	0.72	2.1.1.146	Isoeugenol	Phenylpropanoid	*Ocimum basilicum* (basil)		Not tested
6	0.69	2.1.1.38	O-Demethylpuromycin	Antibiotic	*Streptomyces alboniger*	SaPurOMT	No activity
7	0.68	2.1.1.6	Catechol	Phenol	Diverse		Not tested
8	0.66	2.1.1.212	2,4',7-Trihydroxy-isoflavanone	Flavanonoid	*Lotus japonica*	LjFlaOMT	No activity
9	0.64	2.1.1.4	N-Acetylserotonin	Neurotransmitter	*Homo sapiens* (human)		Not tested
10	**0.64**	**2.1.1.117**	***(S)*-Scoulerine**	**BIA**	***P. somniferum***	**yPsS9OMT**	**Already in pathway, active**
11	0.63	2.1.1.231	4'-Hydroxyflavone	Flavonoid	*Glycine max* (soybean)		Not tested
12	0.63	2.1.1.68	(*E*)-Caffeate	Phenylpropanoid	Diverse		Not tested
13	0.62	2.1.1.104	Caffeoyl-CoA	Phenylpropanoid	*Arabidopsis thaliana*	AtCafOMT	No activity

14	0.61	2.1.1.150	(*E*)-Caffeate	Phenylpropanoid	*Medicago sativa* (alfalfa)	Not tested
15	0.61	2.1.1.222	3-Demethylubiquinol	Quinone	Diverse bacteria	Not tested
16	0.60	2.1.1.279	*trans*-Anol	Phenol	*Pimpinella anisum* (anise)	Not tested
17	0.60	2.1.1.94	16-Hydroxytabersonine	Terpene indole alkaloid	*Catharanthus roseus*	Not tested
18	0.59	2.1.1.114	3,4-Dihydroxy-5-*all*-*trans*-polyprenylbenzoate	Quinone	Diverse, incl. *S. cerevisiae*	Natively present in yeast

The top enzyme classes yielded promising candidates for *in vivo* testing. The first candidate, reticuline 7-*O*-methyltransferase (EC 2.1.1.291), has a BridgIT score of 0.98, making it a good candidate for *in vivo* testing; one variant occurs in *Papaver somniferum*. Ranked second (BridgIT score of 0.76) is the enzyme columbamine *O*-methyltransferase (EC 2.1.1.118; variant from *Coptis japonica* referred to here as CjColOMT), which converts *(S)*-columbamine to *(S)*-palmatine, a similar reaction to our target reaction. A literature search showed that CjColOMT has previously been found to exhibit promiscuous activity *in vitro* on *(S)*-tetrahydrocolumbamine[100]. However, while KEGG catalogues the methylation of (*S*)-tetrahydrocolumbamine to produce (*S*)-tetrahydropalmatine, it does not link it to *CjColOMT* or any other known gene sequence.

The analysis further showed that the *O*-methyltransferases (OMTs) in the noscapine pathway are among the top-ranked candidates for catalyzing the predicted reaction. It has been shown that the majority of metabolic reactions are catalyzed by promiscuous enzymes[101], and enzymes that participate in specialized metabolism are even more likely to be promiscuous[102]–[104]. The potential promiscuity of the noscapine biosynthetic enzymes is thus unsurprising, especially if promiscuous activity is seen on other pathway intermediates that necessarily resemble their native substrates structurally. The enzymes 6OMT (EC 2.1.1.128) and 4'OMT (EC 2.1.1.116), which *O*-methylate the noscapine pathway intermediates *(S)*-norcoclaurine and *(S)*-3'-hydroxy-*N*-methylcoclaurine, respectively, are ranked third and fourth, with BridgIT scores of 0.75. The enzyme S9OMT (EC 2.1.1.117), is ranked tenth with a BridgIT score of 0.64. The high BridgIT scores associated with these three enzymes indicate their potential for promiscuous activity on *(S)*-tetrahydrocolumbamine. As variants of these three enzymes are already present in the noscapine pathway prior to *(S)*-tetrahydrocolumbamine, their potential to produce *(S)*-tetrahydropalmatine will necessarily be evaluated *in vivo*.

5.3.3.5 Two predicted enzymes enable tetrahydropalmatine production *in vitro* and *in vivo*

The preceding workflow generates a ranked list of candidate enzymes predicted to produce the target product. Validation of candidate enzymes can be performed *in vitro* and/or *in vivo* in the context of a heterologous pathway. The ranking of potential enzymes enables a smaller set of enzymes to be tested experimentally, thereby maximizing the success of the project.

We selected seven of the top 18 hits from BridgIT for experimental validation. As described above, three of these enzymes – Ps6OMT, Ps4'OMT, and PsS9OMT – are already present in the biosynthetic pathway upstream of *(S)*-tetrahydrocolumbamine. The other four enzymes were selected based on the diversity of their native substrates, which span a range of less than 300 Da (2,4',7-trihydroxyisoflavanone) to greater than 900 Da (caffeoyl-CoA) (Table 5.5). These four candidate enzymes – columbamine OMT from *Coptis japonica* (CjColOMT, ranked second), *O*-demethylpuromycin OMT from *Streptomyces alboniger* (SaPurOMT, ranked 9th), 2,4',7-Trihydroxyisoflavanone OMT from *Lotus japonica* (LjFlaOMT, ranked 11th), and caffeoyl-coenzyme A OMT from *Arabidopsis thaliana* (AtCafOMT, ranked 17th) – were codon-optimized for expression in *S. cerevisiae*, cloned into high-copy plasmids, and transformed into a *de novo (S)*-tetrahydrocolumbamine producing *S. cerevisiae* strain. *(S)*-Tetrahydropalmatine was produced in every strain tested (Figure 5.7 panel b). However, the strain expressing the highest ranked candidate of those tested, CjColOMT, produced eight-fold more *(S)*-tetrahydropalmatine relative to an empty plasmid control. We hypothesized that the background *(S)*-tetrahydropalmatine in all strains was due to one or more of the other methyltransferases present in the heterologous *(S)*-tetrahydrocolumbamine-producing strain. As these enzymes' native substrates are precursors of, and structurally similar to, *(S)*-tetrahydrocolumbamine, they may possess promiscuous activity on *(S)*-tetrahydrocolumbamine itself. In fact, the other four pathway methyltransferases – S9OMT (acts natively on *(S)*-scoulerine), CNMT (acts natively on coclaurine), 6OMT (acts natively on norcoclaurine), and 4'OMT (acts natively on 6-methyl-*(S)*-laudanosoline) – were assigned high scores by BridgIT for their potential activity on *(S)*-tetrahydrocolumbamine, further supporting this hypothesis.

We next tested each pathway methyltransferase *in vitro* to determine their contribution to the background *(S)*-tetrahydropalmatine production. In the originally constructed heterologous *(S)*-tetrahydrocolumbamine pathway, the four methyltransferases were derived from *Papaver somniferum*, and thus were named Ps6OMT, PsCNMT, Ps4'OMT, and yPsS9OMT (the y prefix on the lattermost denotes that it has been codon-optimized for expression in the yeast *S. cerevisiae*). Ps6OMT, PsCNMT, yPsS9OMT, and CjColOMT expressed well in *E. coli*, but no conditions tested afforded soluble Ps4'OMT. Accordingly, we examined 4'OMT variants from other species and codon-optimized three for expression in *E. coli* – Cj4'OMT from *Coptis japonica*, Ec4'OMT from *Eschscholzia californica*, and Tf4'OMT from *Thalictrum flavum*. These variants expressed well in *E. coli* and were purified for *in vitro* analysis. As these 4'OMT variants might not possess the same substrate promiscuity as the variant originally tested (Ps4'OMT), we created strains with Ps4'OMT replaced with each alternative 4'OMT codon-optimized for expression in *S. cerevisiae*. We verified that, in each of these strains, *(S)*-tetrahydropalmatine was still observed and that expression of CjColOMT resulted in 3- to 7-fold increased production of *(S)*-tetrahydropalmatine (Figure 5.7 panel c).

We tested each pathway methyltransferase and CjColOMT *in vitro* to determine which convert *(S)*-tetrahydrocolumbamine to *(S)*-tetrahydropalmatine. *In vitro* reactions were performed with Ps6OMT, PsCNMT, Cj4'OMT, Ec4'OMT, Tf4'OMT, yPsS9OMT, and CjColOMT. Ps6OMT, PsCNMT, and the 4'OMT variants produced no *(S)*-tetrahydropalmatine *in vitro* (Figure 5.7 panel d). While PsCNMT does accept *(S)*-tetrahydrocolumbamine as a substrate, the product is presumably the *N*-methylated derivative, as the mass is consistent with a second methylation event, no *(S)*-tetrahydropalmatine production was observed, and the *N*-position is the only other available site likely to be methylated by a methyltransferase. Of the pathway enzymes, only *Ps*S9OMT produced *(S)*-tetrahydropalmatine *in vitro* and thus is likely the sole source of the background *(S)*-tetrahydropalmatine observed *in vivo*. To further support this hypothesis, a strain lacking both *yPsS9OMT* and *CjColOMT* produced no *(S)*-tetrahydropalmatine (Figure 5.7 panel b). When tested *in vitro*, CjColOMT afforded over 11-fold higher conversion of *(S)*-tetrahydrocolumbamine to *(S)*-tetrahydropalmatine than yPsS9OMT (Figure 5.7 panel d), which is consistent with the significantly higher production of *(S)*-tetrahydropalmatine *in vivo* upon expression of CjColOMT (Figure 5.7 panel b).

5.3.4 Discussion

In silico tools for novel biosynthetic pathway design can guide and accelerate metabolic engineering to produce molecules of interest. In this work, we employed the biochemical reaction prediction tool BNICE.ch[61] to explore potential biosynthesis targets that can be produced from the noscapine pathway. While multiple pathway prediction tools have been reported, most extract reaction rules automatically from biochemical databases[64], [66], [67], [105], risking the propagation of errors (e.g., unbalanced, orphan or hypothetical multistep reactions) from database entries to the rules. In contrast, BNICE.ch rules are created manually to ensure that the predicted reactions follow biochemical logic. Furthermore, typical retrobiosynthetic approaches focus on a single predetermined compound, whereas our workflow quickly identifies a large number of candidate molecules without requiring prior knowledge of their identities. The high number of available tools stands in contrast to the small number of reported experimental validations of novel, predicted reactions. The first successfully predicted novel bioproduction pathway was established for 1,4-butanediol[106] using the commercial tool SimPheny which, like BNICE.ch, relies on expert-curated generalized reaction rules. Furthermore, novel reactions predicted by BNICE.ch in the ATLAS of Biochemistry[107], [108], a repository of hypothetical biochemical reactions, have only recently been experimentally tested and validated[109]. Both of the examples of successful implementation of predicted novel reactions to date have utilized expert-curated reaction rules.

Once a pathway has been designed, enzymes need to be found to catalyze the predicted biotransformations. Available tools for enzyme function prediction determine the structural similarity of the desired reaction's reactants and products to substrate and products of known enzymes[4]–[7]. In contrast to other tools, BridgIT incorporates information encoded in the BNICE.ch reaction rules to identify the reactive site and then

examines the atom connectivity around the reactive sites of the known and desired substrates. While all mentioned tools benchmarked their predictive capacity on datasets of known enzyme-reaction pairings, no direct experimental validation of an enzyme prediction tool has been reported to our knowledge.

In this study, BNICE.ch identified 15 potential compounds that are one reaction step from an intermediate of the noscapine biosynthetic pathway. We chose to rank these compounds by the sum of their reports in the scientific literature and patents, in order to identify compounds of known biological interest. Our workflow can utilize other ranking algorithms; for example, if searching for new drug candidates, Lipinski's rule of five[110] could be employed, prioritizing compounds over a given molecular mass, calculated partition coefficient, and/or number of hydrogen bond donors and acceptors. One could also prioritize the potential compounds' chemical novelty in order to most effectively leverage the biosynthesis platform to manufacture molecules that cannot be synthesized chemically.

The top two compounds in our ranking that are one biosynthetic step from a noscapine pathway intermediate were berberine and (*S*)-tetrahydropalmatine. The heterologous biosynthesis of berberine has been previously reported[111]; however, the final reaction in its biosynthesis in this strain occurs spontaneously, as the enzyme thought to carry out its biosynthesis in plants appears to be inactive in *S. cerevisiae*, as does a related enzyme[112]. We therefore chose to focus our efforts on *(S)*-tetrahydropalmatine, as numerous methyltransferases have been reported to be active in *S. cerevisiae*, thus decreasing the likelihood that we would encounter false negatives due to lack of expression or proper folding. We recently reported the *de novo* heterologous biosynthesis of (*S*)-tetrahydropalmatine in *S. cerevisiae* via an engineered variant of TfS9OMT, a homologue of PsS9OMT from *Thalictrum flavum*[91]. In particular, the biosynthesis of (*S*)-tetrahydropalmatine was observed with one of two native isoforms of TfS9OMT tested at a level of 0.7 µg/L, and was then increased over fivefold via structure-guided engineering, ultimately yielding a titer of 3.60 µg/L. In contrast, using BridgIT we identified a scoulerine 9-O-methyltransferase (PsS9OMT) and a columbamine O-methyltransferase (CjColOMT) that both perform this transformation, and their expression together in *Saccharomyces cerevisiae* led to a titer of 3.45 µg/L using only native, non-engineered enzymes, nearly matching the titer reported with the best engineered TfS9OMT variant. Replacement or supplementation of PsS9OMT with the engineered TfS9OMT variant could increase our titer of (*S*)-tetrahydropalmatine, or active-site mutagenesis, as was performed for TfS9OMT, could enhance the activity of PsS9OMT or CjColOMT.

The ability of CjColOMT and PsS9OMT to methylate (*S*)-tetrahydrocolumbamine may seem unsurprising, as the native substrates of both enzymes are chemically similar to (*S*)-tetrahydrocolumbamine (Figure 5.7 panel a). In fact, both of these enzymes have been reported to have promiscuous activity toward *(S)*-tetrahydrocolumbamine *in vitro*[113]; however, these non-native activities were not available in our reference database (KEGG). While KEGG includes an entry on the conversion of *(S)*-tetrahydrocolumbamine

to *(S)*-tetrahydropalmatine, this is an "orphan" reaction with no gene or protein sequence associated with it. Recent studies have indicated that 40-50% of all reactions catalogued in KEGG are orphan reactions[114], [115]. In some of these cases, non-native activity data may be available, but is buried in literature and not readily accessible via existing databases, and thus might be overlooked by or unavailable to researchers. In such cases, our computational workflow can provide predictions to guide researchers to enzyme candidates to investigate further, both experimentally and in the existing literature. Furthermore, in cases where the desired non-native enzyme activities have not been reported, our workflow has demonstrated the capability to infer likely off-target activity from only native enzyme data.

This work serves as a proof-of-concept that our computational workflow can use a heterologous biosynthetic pathway to identify a series of potential products and the enzymes required to make those products, thus generating a starting point for subsequent optimization. Protein engineering can then be employed to substantially increase the activity of the integrated enzyme, as has been demonstrated for many classes of enzymes in the past[116]–[118]. Recent years have seen a dramatic increase in the complexity of biosynthetic pathways expressed in heterologous hosts[53], as well as in the efficiency with which these pathways have been reconstructed, spurred by advances in DNA synthesis, sequencing, analytical techniques, and methods for genetic engineering. As increasing numbers of heterologous biosynthetic pathways become available to the research community, as they have for such diverse compound classes as noscapinoids[54], opioids[119], flavonoids[120], [121], cannabinoids[122], and carotenoids[123], computational tools to leverage these pathways for the production of additional products of interest will become increasingly useful. As the number of reported enzymes and compounds also increases, reflected by the continuous growth of biochemical databases like KEGG, we anticipate that computational tools will play a vital role in leveraging this vast amount of data to drive engineering efforts towards the bioproduction of valuable chemicals and pharmaceuticals.

5.4 References

[1] D. R. Nielsen and T. S. Moon, "From promise to practice," *EMBO Rep.*, vol. 14, no. 12, pp. 1034–1038, Dec. 2013, doi: 10.1038/embor.2013.178.

[2] M. Sorokina, M. Stam, C. Medigue, O. Lespinet, and D. Vallenet, "Profiling the orphan enzymes," *Biol. Direct*, p. 9, 2014.

[3] S. D. Copley, "Shining a light on enzyme promiscuity," *Curr. Opin. Struct. Biol.*, vol. 47, pp. 167–175, Dec. 2017, doi: 10.1016/j.sbi.2017.11.001.

[4] N. Hadadi, H. MohammadiPeyhani, L. Miskovic, M. Seijo, and V. Hatzimanikatis, “Enzyme annotation for orphan and novel reactions using knowledge of substrate reactive sites.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 116, no. 15, p. 201818877, Mar. 2019, doi: 10.1073/pnas.1818877116.

[5] S. A. Rahman, S. M. Cuesta, N. Furnham, G. L. Holliday, and J. M. Thornton, “EC-BLAST: a tool to automatically search and compare enzyme reactions,” *Nat. Methods*, vol. 11, no. 2, pp. 171–174, Feb. 2014, doi: 10.1038/nmeth.2803.

[6] Y. Yamanishi, M. Hattori, M. Kotera, S. Goto, and M. Kanehisa, “E-zyme: predicting potential EC numbers from the chemical transformation pattern of substrate-product pairs,” *Bioinformatics*, vol. 25, no. 12, pp. i179–i186, Jun. 2009, doi: 10.1093/bioinformatics/btp223.

[7] P. Carbonell *et al.*, “Selenzyme: enzyme selection tool for pathway design,” *Bioinformatics*, vol. 34, no. 12, pp. 2153–2154, Jun. 2018, doi: 10.1093/bioinformatics/bty065.

[8] N. Hadadi, H. MohammadiPeyhani, L. Miskovic, M. Seijo, and V. Hatzimanikatis, “Enzyme annotation for orphan and novel reactions using knowledge of substrate reactive sites,” *Proc. Natl. Acad. Sci.*, p. 201818877, Mar. 2019, doi: 10.1073/pnas.1818877116.

[9] “Adipic Acid Market Size Price | Global Industry Trends Report, 2020.” https://www.grandviewresearch.com/industry-analysis/adipic-acid-market (accessed Feb. 19, 2018).

[10] A. Shimizu, K. Tanaka, and M. Fujimori, “Abatement technologies for N2O emissions in the adipic acid industry,” *Chemosphere - Glob. Change Sci.*, vol. 2, no. 3, pp. 425–434, Jul. 2000, doi: 10.1016/S1465-9972(00)00024-6.

[11] O. US EPA, “Overview of Greenhouse Gases,” *US EPA*, Dec. 23, 2015. https://www.epa.gov/ghgemissions/overview-greenhouse-gases (accessed Nov. 01, 2019).

[12] A. R. Ravishankara, J. S. Daniel, and R. W. Portmann, “Nitrous Oxide (N2O): The Dominant Ozone-Depleting Substance Emitted in the 21st Century,” *Science*, vol. 326, no. 5949, pp. 123–125, Oct. 2009, doi: 10.1126/science.1176985.

[13] “Tertiary N_2O Abatement | Shell Catalysts & Technologies | Shell Global.” https://www.shell.com/business-customers/catalysts-technologies/catalysts/environmental-catalysts/n20-abatement.html.html#iframe=L24yby1hYmF0ZW1lbnQ (accessed Nov. 01, 2019).

[14] “IPCC - Task Force on National Greenhouse Gas Inventories.” https://www.ipcc-nggip.iges.or.jp/public/gp/english/ (accessed Feb. 04, 2020).

[15] M. Faber, “Process for producing adipic acid from biomass,” US4400468A, Aug. 23, 1983.

[16] K. M. Draths and J. W. Frost, “Environmentally compatible synthesis of adipic acid from D-glucose,” *J. Am. Chem. Soc.*, vol. 116, no. 1, pp. 399–400, Jan. 1994, doi: 10.1021/ja00080a057.

[17] M. Kohlstedt *et al.*, “From lignin to nylon: Cascaded chemical and biochemical conversion using metabolically engineered Pseudomonas putida,” *Metab. Eng.*, vol. 47, pp. 279–293, May 2018, doi: 10.1016/j.ymben.2018.03.003.

[18] R. R. Yocum, W. Gong, S. Dole, R. Sillers, M. Gandhi, and J. G. Pero, “Production of muconic acid from genetically engineered microorganisms,” US20150044755A1, Feb. 12, 2015.

[19] G. Wang *et al.*, “Improvement of cis,cis-Muconic Acid Production in Saccharomyces cerevisiae through Biosensor-Aided Genome Engineering,” *ACS Synth. Biol.*, vol. 9, no. 3, pp. 634–646, Mar. 2020, doi: 10.1021/acssynbio.9b00477.

[20] S. Picataggio and T. Beardslee, “Biological methods for preparing adipic acid,” US8343752B2, Jan. 01, 2013.

[21] Y. Deng and Y. Mao, “Production of adipic acid by the native-occurring pathway in *Thermobifida fusca* B6,” *J. Appl. Microbiol.*, vol. 119, no. 4, pp. 1057–1063, Oct. 2015, doi: 10.1111/jam.12905.

[22] S. Cheong, J. M. Clomburg, and R. Gonzalez, “Energy- and carbon-efficient synthesis of functionalized small molecules in bacteria using non-decarboxylative Claisen condensation reactions,” *Nat. Biotechnol.*, vol. 34, no. 5, pp. 556–561, May 2016, doi: 10.1038/nbt.3505.

[23] S. F. Altschul *et al.*, “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs,” *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389–3402, Sep. 1997, doi: 10.1093/nar/25.17.3389.

[24] N. Hadadi, H. MohammadiPeyhani, L. Miskovic, M. Seijo, and V. Hatzimanikatis, “Enzyme annotation for orphan and novel reactions using knowledge of substrate reactive sites,” *Proc. Natl. Acad. Sci.*, vol. 116, no. 15, pp. 7298–7307, Apr. 2019, doi: 10.1073/pnas.1818877116.

[25] E. J. Kerkhoven, K. R. Pomraning, S. E. Baker, and J. Nielsen, “Regulation of amino-acid metabolism controls flux to lipid accumulation in *Yarrowia lipolytica*,” *Npj Syst. Biol. Appl.*, vol. 2, p. 16005, Mar. 2016, doi: 10.1038/npjsba.2016.5.

[26] A. Varma and B. O. Palsson, “Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use,” *Bio/Technology*, vol. 12, no. 10, Art. no. 10, Oct. 1994, doi: 10.1038/nbt1094-994.

[27] C. S. Henry, L. J. Broadbelt, and V. Hatzimanikatis, “Thermodynamics-Based Metabolic Flux Analysis,” *Biophys. J.*, vol. 92, no. 5, pp. 1792–1805, Mar. 2007, doi: 10.1529/biophysj.106.093138.

[28] C. Holkenbrink *et al.*, “EasyCloneYALI: CRISPR/Cas9-Based Synthetic Toolbox for Engineering of the Yeast Yarrowia lipolytica,” *Biotechnol. J.*, vol. 0, no. 0, p. 1700543, doi: 10.1002/biot.201700543.

[29] M. M. Jessop-Fabre *et al.*, “EasyClone-MarkerFree: A vector toolkit for marker-less integration of genes into Saccharomyces cerevisiae via CRISPR-Cas9,” *Biotechnol. J.*, vol. 11, no. 8, pp. 1110–1117, doi: 10.1002/biot.201600147.

[30] N. B. Jensen *et al.*, “EasyClone: method for iterative chromosomal integration of multiple genes in Saccharomyces cerevisiae,” *FEMS Yeast Res.*, vol. 14, no. 2, pp. 238–248, 2014, doi: 10.1111/1567-1364.12118.

[31] S. V. Kamzolova and I. G. Morgunov, “α-Ketoglutaric acid production from rapeseed oil by <Emphasis Type="Italic">Yarrowia lipolytica</Emphasis> yeast,” *Appl. Microbiol. Biotechnol.*, vol. 97, no. 12, pp. 5517–5525, Jun. 2013, doi: 10.1007/s00253-013-4772-6.

[32] M. Tokunaga, Y. Nakano, and S. Kitaoka, “Separation and properties of the NAD-linked and NADP-linked isozymes of succinic semialdehyde dehydrogenase in Euglena gracilis,” *Biochim. Biophys. Acta BBA - Enzymol.*, vol. 429, no. 1, pp. 55–62, Mar. 1976, doi: 10.1016/0005-2744(76)90029-2.

[33] P. Siegert *et al.*, “Exchanging the substrate specificities of pyruvate decarboxylase from Zymomonas mobilis and benzoylformate decarboxylase from Pseudomonas putida,” *Protein Eng. Des. Sel.*, vol. 18, no. 7, pp. 345–357, Jul. 2005, doi: 10.1093/protein/gzi035.

[34] D. Esser *et al.*, “Unraveling the function of paralogs of the aldehyde dehydrogenase super family from <Emphasis Type="Italic">Sulfolobus solfataricus</Emphasis>,” *Extremophiles*, vol. 17, no. 2, pp. 205–216, Mar. 2013, doi: 10.1007/s00792-012-0507-3.

[35] R. M. Drevland, Y. Jia, D. R. J. Palmer, and D. E. Graham, “Methanogen Homoaconitase Catalyzes Both Hydrolyase Reactions in Coenzyme B Biosynthesis,” *J. Biol. Chem.*, vol. 283, no. 43, pp. 28888–28896, Oct. 2008, doi: 10.1074/jbc.M802159200.

[36] J. Dahlin *et al.*, “Multi-Omics Analysis of Fatty Alcohol Production in Engineered Yeasts Saccharomyces cerevisiae and Yarrowia lipolytica,” *Front. Genet.*, vol. 10, 2019, doi: 10.3389/fgene.2019.00747.

[37] L. A. Urrestarazu, C. W. Borell, and J. K. Bhattacharjee, “General and specific controls of lysine biosynthesis in Saccharomyces cerevisiae,” *Curr. Genet.*, vol. 9, no. 5, pp. 341–344, May 1985, doi: 10.1007/BF00421603.

[38] B. Becker, A. Feller, M. E. Alami, E. Dubois, and A. Piérard, “A nonameric core sequence is required upstream of the LYS genes of Saccharomyces cerevisiae for Lys14p-mediated activation and apparent repression by lysine,” *Mol. Microbiol.*, vol. 29, no. 1, pp. 151–163, Jul. 1998, doi: 10.1046/j.1365-2958.1998.00916.x.

[39] A. Feller, F. Ramos, A. Piérard, and E. Dubois, “In Saccharomyces cerevisae, feedback inhibition of homocitrate synthase isoenzymes by lysine modulates the activation of LYS gene expression by Lys14p,” *Eur. J. Biochem.*, vol. 261, no. 1, pp. 163–170, doi: 10.1046/j.1432-1327.1999.00262.x.

[40] T. Esikova, O. Ponamoreva, B. Baskunov, S. Taran, and A. Boronin, “Transformation of low-molecular linear caprolactam oligomers by caprolactam-degrading bacteria,” *J. Chem. Technol. Biotechnol.*, vol. 87, no. 9, pp. 1284–1290, 2012, doi: 10.1002/jctb.3789.

[41] H. Iwaki, Y. Hasegawa, M. Teraoka, T. Tokuyama, H. Bergeron, and P. C. K. Lau, “Identification of a Transcriptional Activator (ChnR) and a 6-Oxohexanoate Dehydrogenase (ChnE) in the Cyclohexanol Catabolic Pathway in Acinetobacter sp. Strain NCIMB 9871 and Localization of the Genes That Encode Them,” *Appl. Environ. Microbiol.*, vol. 65, no. 11, pp. 5158–5162, Nov. 1999.

[42] P. Scarcia, L. Palmieri, G. Agrimi, F. Palmieri, and H. Rottensteiner, “Three mitochondrial transporters of Saccharomyces cerevisiae are essential for ammonium fixation and lysine biosynthesis in synthetic minimal medium,” *Mol. Genet. Metab.*, vol. 122, no. 3, pp. 54–60, Nov. 2017, doi: 10.1016/j.ymgme.2017.07.004.

[43] L. Palmieri, G. Agrimi, M. J. Runswick, I. M. Fearnley, F. Palmieri, and J. E. Walker, “Identification in Saccharomyces cerevisiae of Two Isoforms of a Novel Mitochondrial Transporter for 2-Oxoadipate and 2-Oxoglutarate,” *J. Biol. Chem.*, vol. 276, no. 3, pp. 1916–1922, Jan. 2001, doi: 10.1074/jbc.M004332200.

[44] M. Kiebler, K. Becker, N. Pfanner, and W. Neupert, “Mitochondrial protein import: Specific recognition and membrane translocation of preproteins,” *J. Membr. Biol.*, vol. 135, no. 3, pp. 191–207, Sep. 1993, doi: 10.1007/BF00211091.

[45] R. T. Baker *et al.*, “Organic acids from homocitrate and homocitrate derivatives,” US20170113993A1, Apr. 27, 2017.

[46] N. S. Kruyer and P. Peralta-Yahya, “Metabolic engineering strategies to bio-adipic acid production,” *Curr. Opin. Biotechnol.*, vol. 45, pp. 136–143, Jun. 2017, doi: 10.1016/j.copbio.2017.03.006.

[47] O. G. Chernyavskaya, N. V. Shishkanova, A. P. Il’chenko, and T. V. Finogenova, “Synthesis of α-ketoglutaric acid by Yarrowia lipolytica yeast grown on ethanol,” *Appl. Microbiol. Biotechnol.*, vol. 53, no. 2, pp. 152–158, Feb. 2000, doi: 10.1007/s002530050002.

[48] X. Yin, C. Madzak, G. Du, J. Zhou, and J. Chen, “Enhanced alpha-ketoglutaric acid production in Yarrowia lipolytica WSH-Z06 by regulation of the pyruvate carboxylation pathway,” *Appl. Microbiol. Biotechnol.*, vol. 96, no. 6, pp. 1527–1537, Dec. 2012, doi: 10.1007/s00253-012-4192-z.

[49] J. Zhou, X. Yin, C. Madzak, G. Du, and J. Chen, “Enhanced α-ketoglutarate production in Yarrowia lipolytica WSH-Z06 by alteration of the acetyl-CoA metabolism,” *J. Biotechnol.*, vol. 161, no. 3, pp. 257–264, Oct. 2012, doi: 10.1016/j.jbiotec.2012.05.025.

[50] G. M. Cragg and D. J. Newman, “Natural products: A continuing source of novel drug leads,” *Biochimica et Biophysica Acta - General Subjects*, vol. 1830, no. 6. Elsevier, pp. 3670–3695, Jun. 01, 2013, doi: 10.1016/j.bbagen.2013.02.008.

[51] D. J. Newman and G. M. Cragg, “Natural products as sources of new drugs over the 30 years from 1981 to 2010,” *J. Nat. Prod.*, vol. 75, no. 3, pp. 311–335, Mar. 2012, doi: 10.1021/np200906s.

[52] D.-K. Ro *et al.*, “Production of the antimalarial drug precursor artemisinic acid in engineered yeast,” *Nature*, vol. 440, no. 7086, pp. 940–943, Apr. 2006, doi: 10.1038/nature04640.

[53] A. Cravens, J. Payne, and C. D. Smolke, “Synthetic biology strategies for microbial biosynthesis of plant natural products,” *Nat. Commun.*, vol. 10, no. 1, p. 2142, Dec. 2019, doi: 10.1038/s41467-019-09848-w.

[54] Y. Li, S. Li, K. Thodey, I. Trenchard, A. Cravens, and C. D. Smolke, “Complete biosynthesis of noscapine and halogenated alkaloids in yeast,” *Proc. Natl. Acad. Sci.*, vol. 115, no. 17, pp. E3922–E3931, Apr. 2018, doi: 10.1073/PNAS.1721469115.

[55] K. Ye *et al.*, “Opium alkaloid noscapine is an antitumor agent that arrests metaphase and induces apoptosis in dividing cells,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 95, no. 4, pp. 1601–1606, Feb. 1998, doi: 10.1073/pnas.95.4.1601.

[56] B. Israel, J. Geller, and M. Rogosnitzky, “Noscapine Inhibits Human Prostate Cancer Progression and Metastasis in a Mouse Model,” *Anticancer Res.*, vol. 28, pp. 3701–3704, 2008.

[57] N. P. Joshi HC, Salil A, Bughani U, “Noscapinoids: a new class of anticancer drugs demand bio-technological intervention,” in *Medicinal Plant Biotechnology*, R. Arora, Ed. CAB e-Books.

[58] N. Hadadi and V. Hatzimanikatis, “Design of computational retrobiosynthesis tools for the design of de novo synthetic pathways,” *Curr. Opin. Chem. Biol.*, vol. 28, pp. 99–104, Oct. 2015, doi: 10.1016/J.CBPA.2015.06.025.

[59] J. G. Jeffryes, S. M. D. Seaver, J. P. Faria, and C. S. Henry, “A pathway for every product? Tools to discover and design plant metabolism,” *Plant Sci.*, Mar. 2018, doi: 10.1016/J.PLANTSCI.2018.03.025.

[60] G.-M. M. Lin, R. Warden-Rothman, and C. A. Voigt, “Retrosynthetic design of metabolic pathways to chemicals not found in nature,” *Curr. Opin. Syst. Biol.*, vol. 14, pp. 82–107, Apr. 2019, doi: 10.1016/J.COISB.2019.04.004.

[61] V. Hatzimanikatis, C. Li, J. A. Ionita, C. S. Henry, M. D. Jankowski, and L. J. Broadbelt, “Exploring the diversity of complex metabolic networks,” *Bioinformatics*, vol. 21, no. 8, pp. 1603–1609, Apr. 2005, doi: 10.1093/bioinformatics/bti213.

[62] J. Wicker *et al.*, “enviPath – The environmental contaminant biotransformation pathway resource,” *Nucleic Acids Res.*, p. gkv1229, Nov. 2015, doi: 10.1093/nar/gkv1229.

[63] M. A. Campodonico, B. A. Andrews, J. A. Asenjo, B. O. Palsson, and A. M. Feist, “Generation of an atlas for commodity chemical production in Escherichia coli and a novel pathway prediction algorithm, GEM-Path,” *Metab. Eng.*, vol. 25, pp. 140–158, Sep. 2014, doi: 10.1016/J.YMBEN.2014.07.009.

[64] S. Ding *et al.*, “novoPathFinder: a webserver of designing novel-pathway with integrating GEM-model,” *Nucleic Acids Res.*, no. 1, 2020, doi: 10.1093/nar/gkaa230.

[65] A. Kumar, L. Wang, C. Y. Ng, and C. D. Maranas, “Pathway design using de novo steps through uncharted biochemical spaces,” *Nat. Commun.*, vol. 9, no. 1, p. 184, Jan. 2018, doi: 10.1038/s41467-017-02362-x.

[66] T. V. Sivakumar, V. Giri, J. H. Park, T. Y. Kim, and A. Bhaduri, “ReactPRED: A tool to predict and analyze biochemical reactions,” *Bioinformatics*, p. btw491, Aug. 2016, doi: 10.1093/bioinformatics/btw491.

[67] B. Delépine, T. Duigou, P. Carbonell, and J.-L. Faulon, “RetroPath2.0: A retrosynthesis workflow for metabolic engineers,” *Metab. Eng.*, vol. 45, pp. 158–170, Jan. 2018, doi: 10.1016/j.ymben.2017.12.002.

[68] J. D. Tyzack, A. J. M. Ribeiro, N. Borkakoti, and J. M. Thornton, “Exploring Chemical Biosynthetic Design Space with Transform-MinER,” *ACS Synth. Biol.*, vol. 8, no. 11, pp. 2494–2506, Nov. 2019, doi: 10.1021/acssynbio.9b00105.

[69] M. Tokić *et al.*, “Discovery and evaluation of biosynthetic pathways for the production of five methyl ethyl ketone precursors,” *ACS Synth. Biol.*, p. acssynbio.8b00049, Jul. 2018, doi: 10.1021/acssynbio.8b00049.

[70] M. Moura, J. Finkle, S. Stainbrook, J. Greene, L. J. Broadbelt, and K. E. J. Tyo, “Evaluating enzymatic synthesis of small molecule drugs,” *Metab. Eng.*, vol. 33, pp. 138–147, Jan. 2016, doi: 10.1016/J.YMBEN.2015.11.006.

[71] L. Wang, C. Y. Ng, S. Dash, and C. D. Maranas, “Exploring the combinatorial space of complete pathways to chemicals.,” *Biochem. Soc. Trans.*, vol. 46, no. 3, pp. 513–522, Jun. 2018, doi: 10.1042/BST20170272.

[72] M. Lin, F. Chueh, M. Hsieh, and C. Chen, “ANTIHYPERTENSIVE EFFECTS OF dl-TETRAHYDROPALMATINE: AN ACTIVE PRINCIPLE ISOLATED FROM CORYDALIS,” *Clin. Exp. Pharmacol. Physiol.*, vol. 23, no. 8, pp. 738–745, Aug. 1996, doi: 10.1111/j.1440-1681.1996.tb01769.x.

[73] W. Chung Leung, H. Zheng, M. Huen, S. Lun Law, and H. Xue, “Anxiolytic-like action of orally administered dl-tetrahydropalmatine in elevated plus-maze,” *Prog. Neuropsychopharmacol. Biol. Psychiatry*, vol. 27, no. 5, pp. 775–779, Aug. 2003, doi: 10.1016/S0278-5846(03)00108-8.

[74] J. R. Mantsch *et al.*, “Levo-tetrahydropalmatine attenuates cocaine self-administration and cocaine-induced reinstatement in rats,” *Psychopharmacology (Berl.)*, vol. 192, no. 4, pp. 581–591, May 2007, doi: 10.1007/s00213-007-0754-7.

[75] V. Hatzimanikatis, C. Li, J. A. Ionita, C. S. Henry, M. D. Jankowski, and L. J. Broadbelt, “Exploring the diversity of complex metabolic networks,” *Bioinformatics*, vol. 21, no. 8, pp. 1603–1609, Apr. 2005, doi: 10.1093/bioinformatics/bti213.

[76] M. Kanehisa and S. Goto, “KEGG: kyoto encyclopedia of genes and genomes,” *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, Jan. 2000.

[77] R. Overbeek *et al.*, “The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes,” *Nucleic Acids Res.*, vol. 33, no. 17, pp. 5691–5702, Sep. 2005, doi: 10.1093/nar/gki866.

[78] D. S. Wishart *et al.*, “HMDB: the Human Metabolome Database.,” *Nucleic Acids Res.*, vol. 35, no. Database issue, pp. D521-6, Jan. 2007, doi: 10.1093/nar/gkl923.

[79] R. Caspi *et al.*, “The MetaCyc database of metabolic pathways and enzymes,” *Nucleic Acids Res.*, vol. 46, no. D1, pp. D633–D639, Jan. 2018, doi: 10.1093/nar/gkx935.

[80] L. Jeske, S. Placzek, I. Schomburg, A. Chang, and D. Schomburg, “BRENDA in 2019: a European ELIXIR core data resource,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D542–D549, Jan. 2019, doi: 10.1093/nar/gky1048.

[81] S. Moretti, O. Martin, T. Van Du Tran, A. Bridge, A. Morgat, and M. Pagni, “MetaNetX/MNXref – reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks,” *Nucleic Acids Res.*, vol. 44, no. D1, pp. D523–D526, Jan. 2016, doi: 10.1093/nar/gkv1117.

[82] A. Morgat *et al.*, "Updates in Rhea--a manually curated resource of biochemical reactions.," *Nucleic Acids Res.*, vol. 43, no. Database issue, pp. D459-64, Jan. 2015, doi: 10.1093/nar/gku961.

[83] J. Schellenberger, J. O. Park, T. M. Conrad, and B. Ø. Palsson, "BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions," *BMC Bioinformatics*, vol. 11, no. 1, p. 213, Apr. 2010, doi: 10.1186/1471-2105-11-213.

[84] P. Schläpfer *et al.*, "Genome-Wide Prediction of Metabolic Enzymes, Pathways, and Gene Clusters in Plants," *Plant Physiol.*, vol. 173, no. 4, pp. 2041–2059, Apr. 2017, doi: 10.1104/PP.16.01942.

[85] F. M. Afendi *et al.*, "KNApSAcK Family Databases: Integrated Metabolite–Plant Species Databases for Multifaceted Plant Research," *Plant Cell Physiol.*, vol. 53, no. 2, pp. e1–e1, Feb. 2012, doi: 10.1093/pcp/pcr165.

[86] J. Hastings *et al.*, "ChEBI in 2016: Improved services and an expanding collection of metabolites," *Nucleic Acids Res.*, vol. 44, no. Database issue, p. D1214, Jan. 2016, doi: 10.1093/NAR/GKV1031.

[87] A. Gaulton *et al.*, "The ChEMBL database in 2017," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D945–D954, Jan. 2017, doi: 10.1093/nar/gkw1074.

[88] S. Kim *et al.*, "PubChem 2019 update: improved access to chemical data," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D1102–D1109, Jan. 2019, doi: 10.1093/nar/gky1033.

[89] S. Kim, P. A. Thiessen, T. Cheng, B. Yu, and E. E. Bolton, "An update on PUG-REST: RESTful interface for programmatic access to PubChem.," *Nucleic Acids Res.*, vol. 46, no. W1, pp. W563–W570, Jul. 2018, doi: 10.1093/nar/gky294.

[90] E. Sayers, "A General Introduction to the E-utilities," 2010.

[91] T. R. Valentic, J. T. Payne, and C. D. Smolke, "Structure-Guided Engineering of a Scoulerine 9- *O* -Methyltransferase Enables the Biosynthesis of Tetrahydropalmatrubine and Tetrahydropalmatine in Yeast," *ACS Catal.*, pp. 4497–4509, Mar. 2020, doi: 10.1021/acscatal.9b05417.

[92] O. W. Ryan *et al.*, "Selection of chromosomal DNA libraries using a multiplex CRISPR system," *eLife*, vol. 3, no. August2014, pp. 1–15, Aug. 2014, doi: 10.7554/eLife.03703.

[93] P. Srinivasan and C. D. Smolke, "Engineering a microbial biosynthesis platform for de novo production of tropane alkaloids," *Nat. Commun.*, vol. 10, no. 1, Dec. 2019, doi: 10.1038/s41467-019-11588-w.

[94] S. Galanie, K. Thodey, I. J. Trenchard, M. F. Interrante, and C. D. Smolke, "Complete biosynthesis of opioids in yeast," *Science*, vol. 349, no. 6252, pp. 1095–1100, Sep. 2015, doi: 10.1126/science.aac9373.

[95] I. Schomburg, A. Chang, O. Hofmann, C. Ebeling, F. Ehrentreich, and D. Schomburg, “BRENDA: a resource for enzyme data and metabolic information,” *Trends Biochem. Sci.*, vol. 27, no. 1, pp. 54–56, Jan. 2002, doi: 10.1016/S0968-0004(01)02027-8.

[96] Z. A. King *et al.*, “BiGG Models: A platform for integrating, standardizing and sharing genome-scale models,” *Nucleic Acids Res.*, vol. 44, no. D1, pp. D515–D522, Jan. 2016, doi: 10.1093/nar/gkv1049.

[97] G. A. Johnston, “Advantages of an antagonist: bicuculline and other GABA antagonists,” *Br. J. Pharmacol.*, vol. 169, no. 2, pp. 328–336, May 2013, doi: 10.1111/bph.12127.

[98] K. M. Hawkins and C. D. Smolke, “Production of benzylisoquinoline alkaloids in Saccharomyces cerevisiae,” *Nat. Chem. Biol.*, vol. 4, no. 9, pp. 564–573, Sep. 2008, doi: 10.1038/nchembio.105.

[99] C. Desgrouas *et al.*, “Ethnobotany, phytochemistry and pharmacology of Stephania rotunda Lour,” *Journal of Ethnopharmacology*, vol. 154, no. 3. Elsevier Ireland Ltd, pp. 537–563, Jul. 03, 2014, doi: 10.1016/j.jep.2014.04.024.

[100] T. Morishige, E. Dubouzet, K.-B. Choi, K. Yazaki, and F. Sato, “Molecular cloning of columbamine O-methyltransferase from cultured Coptis japonica cells,” *Eur. J. Biochem.*, vol. 269, no. 22, pp. 5659–5667, Nov. 2002, doi: 10.1046/j.1432-1033.2002.03275.x.

[101] H. Nam *et al.*, “Network context and selection in the evolution to enzyme specificity.,” *Science*, vol. 337, no. 6098, pp. 1101–4, Aug. 2012, doi: 10.1126/science.1216861.

[102] A. Babtie, N. Tokuriki, and F. Hollfelder, “What makes an enzyme promiscuous?,” *Current Opinion in Chemical Biology*, vol. 14, no. 2. Elsevier Current Trends, pp. 200–207, Apr. 01, 2010, doi: 10.1016/j.cbpa.2009.11.028.

[103] A. Bar-Even *et al.*, “The moderately efficient enzyme: Evolutionary and physicochemical trends shaping enzyme parameters,” *Biochemistry*, vol. 50, no. 21, pp. 4402–4410, May 2011, doi: 10.1021/bi2002289.

[104] A. Bar-Even, R. Milo, E. Noor, and D. S. Tawfik, “The Moderately Efficient Enzyme: Futile Encounters and Enzyme Floppiness,” *Biochemistry*, vol. 54, no. 32, pp. 4969–4977, Jul. 2015, doi: 10.1021/acs.biochem.5b00621.

[105] A. Kumar, L. Wang, C. Y. Ng, and C. D. Maranas, “Pathway design using de novo steps through uncharted biochemical spaces,” *Nat. Commun.*, vol. 9, no. 1, p. 184, Jan. 2018, doi: 10.1038/s41467-017-02362-x.

[106] H. Yim *et al.*, “Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol,” *Nat. Chem. Biol.*, vol. 7, no. 7, pp. 445–452, Jul. 2011, doi: 10.1038/nchembio.580.

[107] N. Hadadi, J. Hafner, A. Shajkofci, A. Zisaki, and V. Hatzimanikatis, “ATLAS of Biochemistry: A repository of all possible biochemical reactions for synthetic biology and metabolic engineering studies,” *ACS Synth. Biol.*, Jul. 2016, doi: 10.1021/acssynbio.6b00054.

[108] J. Hafner, H. MohammadiPeyhani, A. Sveshnikova, A. Scheidegger, and V. Hatzimanikatis, “Updated ATLAS of Biochemistry with New Metabolites and Improved Enzyme Prediction Power,” *ACS Synth. Biol.*, vol. 9, no. 6, pp. 1479–1482, May 2020, doi: 10.1021/acssynbio.0c00052.

[109] X. Yang *et al.*, “Systematic design and in vitro validation of novel one-carbon assimilation pathways,” *Metab. Eng.*, vol. 56, pp. 142–153, Dec. 2019, doi: 10.1016/J.YMBEN.2019.09.001.

[110] C. A. Lipinski, “Lead- and drug-like compounds: the rule-of-five revolution,” *Drug Discov. Today Technol.*, vol. 1, no. 4, pp. 337–341, Dec. 2004, doi: 10.1016/J.DDTEC.2004.11.007.

[111] S. Galanie and C. D. Smolke, “Optimization of yeast-based production of medicinal protoberberine alkaloids,” *Microb. Cell Factories*, vol. 14, no. 1, p. 144, Sep. 2015, doi: 10.1186/s12934-015-0332-3.

[112] J. M. Hagel, G. A. W. Beaudoin, E. Fossati, A. Ekins, V. J. J. Martin, and P. J. Facchini, “Characterization of a flavoprotein oxidase from opium poppy catalyzing the final steps in sanguinarine and papaverine biosynthesis,” *J. Biol. Chem.*, vol. 287, no. 51, pp. 42972–42983, Dec. 2012, doi: 10.1074/jbc.M112.420414.

[113] T.-T. T. Dang and P. J. Facchini, “Characterization of three O-methyltransferases involved in noscapine biosynthesis in opium poppy.,” *Plant Physiol.*, vol. 159, no. 2, pp. 618–31, Jun. 2012, doi: 10.1104/pp.112.194886.

[114] M. Sorokina, M. Stam, C. Médigue, O. Lespinet, and D. Vallenet, “Profiling the orphan enzymes,” *Biology Direct*, vol. 9, no. 1. BioMed Central Ltd., p. 10, Jun. 06, 2014, doi: 10.1186/1745-6150-9-10.

[115] A. G. Shearer, T. Altman, and C. D. Rhee, “Finding sequences for over 270 orphan enzymes,” *PLoS ONE*, vol. 9, no. 5, May 2014, doi: 10.1371/journal.pone.0097250.

[116] R. Fasan, M. M. Chen, N. C. Crook, and F. H. Arnold, “Engineered alkane-hydroxylating cytochrome P450BM3 exhibiting nativelike catalytic properties,” *Angew. Chem. - Int. Ed.*, vol. 46, no. 44, pp. 8414–8418, Nov. 2007, doi: 10.1002/anie.200702616.

[117] J. T. Payne, C. B. Poor, and J. C. Lewis, “Directed Evolution of RebH for Site-Selective Halogenation of Large Biologically Active Molecules,” *Angew. Chem. Int. Ed.*, vol. 54, no. 14, pp. 4226–4230, Mar. 2015, doi: 10.1002/anie.201411901.

[118] C. K. Savile *et al.*, "Biocatalytic asymmetric synthesis of chiral amines from ketones applied to sitagliptin manufacture," *Science*, vol. 329, no. 5989, pp. 305–309, Jul. 2010, doi: 10.1126/science.1188934.

[119] S. Galanie, K. Thodey, I. J. Trenchard, M. F. Interrante, and C. D. Smolke, "Complete biosynthesis of opioids in yeast," *Science*, vol. 349, no. 6252, pp. 1095–1100, Sep. 2015, doi: 10.1126/science.aac9373.

[120] J. Li, C. Tian, Y. Xia, I. Mutanda, K. Wang, and Y. Wang, "Production of plant-specific flavones baicalein and scutellarein in an engineered E. coli from available phenylalanine and tyrosine," *Metab. Eng.*, vol. 52, pp. 124–133, Mar. 2019, doi: 10.1016/J.YMBEN.2018.11.008.

[121] J. A. Jones *et al.*, "Experimental and computational optimization of an Escherichia coli co-culture for the efficient production of flavonoids," *Metab. Eng.*, vol. 35, pp. 55–63, May 2016, doi: 10.1016/J.YMBEN.2016.01.006.

[122] X. Luo *et al.*, "Complete biosynthesis of cannabinoids and their unnatural analogues in yeast," *Nature*, p. 1, Feb. 2019, doi: 10.1038/s41586-019-0978-9.

[123] C. Zhang, X. Chen, N. D. Lindley, and H.-P. Too, "A 'plug-n-play' modular metabolic system for the production of apocarotenoids," *Biotechnol. Bioeng.*, vol. 115, no. 1, pp. 174–183, Jan. 2018, doi: 10.1002/bit.26462.

Chapter 6 NICEdrug.ch: a workflow for rational drug design and systems-level analysis of drug metabolism

"Medicine heals doubts as well as diseases"

D Karl Marx

The results presented in this chapter have been obtained in collaboration with several people with the author of this thesis as the leading scientist. Kiandokht Haddadi, contributed in collecting and curating drug molecules from different databases, predicting metabolic reactions using BNICE.ch, and similarity evaluating of drug molecules under the supervision of the author. Dr. Anush Chiapino-Pepe presented essential human and malaria parasite genes, performed data analysis and contributed in visualization and investigation. Dr. Jasmin Hafner designed the web interface of NICEdrug.ch and Dr. Noushin Hadadi contributed in conceptualization and methodology development. The author of this thesis designed the whole workflow, studied its application in different case studies and contributed in data analysis, visualization and investigation. Prof. Vassily Hatzimanikatis supervised the project as well as the completion of the manuscript. This chapter is submitted as a manuscript for the publication. The tool developed here, named NICEdrug.ch, is currently hosted on the LCSB sharing platform.

Full list of authors in this paper: H. MohammadiPeyhani, A. Chiappino-Pepe†, K. Haddadi†, J. Hafner, N. Hadadi, and V. Hatzi-manikatis, "Database for drug metabolism and comparisons, NICEdrug.ch, aids discovery and de-sign," bioRxiv, p. 2020.05.28.120782, Jan. 2020, doi: 10.1101/2020.05.28.120782 (under review, † contributed equally, * corresponding author).*

6.1 Introduction

Discovering new non-toxic drugs is required to treat diseases and infections, target drug resistance, and develop personalized treatments. However, identifying, testing, and approving a single small molecule can take decades and billions of dollars—and there is still a high risk that the proposed drug candidate fails. There is an urgent need to define strategies that accelerate the discovery of new, safe, and effective drugs, and the computational screening of *all* possible targets and molecules can help toward this aim. All computational approaches to date have focused on molecular structures without considering the *reactivity* of the molecules in a cell. However, reactivity information and drug metabolism determine which enzymes the drugs will target, the drug's metabolic fate or degradation, and the potential source of its toxicity and side effects. Understanding drug effects in the context of cellular metabolism also offers great promise in evaluating the

reactivity of a new small molecule, the druggability of an enzyme, and the possibility of drug repurposing. Yet, the *in silico* mechanistic analysis of drug biochemistry is relatively unexplored, and no large-scale studies of drug biochemistry in cells have ever been performed.

To systematically illuminate the metabolism and all enzymatic targets of known drugs and hypothetical prodrugs, we have herein performed the first large-scale computational analysis of drug biochemistry and toxicity in the context of human metabolism. To do this, we employed proven tools for analyzing the neighboring atoms around enzyme reactive sites (BridgIT and BNICE.ch). The analysis involved over 250,000 small molecules, and was a huge technical effort spanning the curation and computation of bio- and physico-chemical drug properties. We assembled this in an open-source workflow, NICEdrug.ch, that can generate drug metabolic reports and can be easily accessed and used by researchers, clinicians, and industry partners.

6.1.1 Drug discovery : An ongoing challenge

To assure effective therapies for previously untreated illness, emerging diseases, and personalized medicine, new small molecules are always needed. However, the process to develop new drugs is complex, costly, and time consuming. This is especially problematic considering about 90% of drug candidates in clinical trials are discarded due to unexpected toxicity or other secondary effects. This inefficiency threatens our health care system and economy [1]. Improving how we discover and design new drugs could reduce the time and costs involved in the developmental pipeline and hence is of primary importance to define efficient medical therapies.

6.1.2 How drugs are designed and developed

Current drug discovery techniques often involve high-throughput screens with candidates and a set of target enzymes presumably involved in a disease, which leads to the selection for those candidates with the preferred activity. However, the biochemical space of small molecules and possible targets in the cell is huge, which limits the possible experimental testing. Computational methods for drug pre-screening and discovery are therefore promising. *In silico,* one can systematically search the maximum biochemical space for targets and molecules with desired structures and functions to narrow down the molecules to test experimentally.

There are two main *in silico* strategies for drug discovery: a data-driven approach based on machine learning, or a mechanistic approach based on the available biochemical knowledge. Machine learning (ML) has been successfully used in all stages of drug discovery, from the prediction of targets to the discovery of drug candidates, as shown in some recent studies [2]–[4]. However, ML approaches require big, high-quality data sets of drug activity and associated physiology [4], which might be challenging to obtain when studying drug action mechanisms and side effects in humans. ML also uses trained neural networks, which can lack

interpretability and repeatability. This can make it difficult to explain why the neural networks has chosen a specific result, why it unexpectedly failed for an unseen dataset, and the final results may vary [4].

Mechanistic-based approaches can also rationally identify small molecules in a desired system and do not require such large amounts of data. Such methods commonly screen based on structural similarity to a native enzyme substrate (antimetabolite) or to a known drug (for drug repurposing), considering the complete structure of a molecule to extract information about protein-ligand fitness [5], [6]. However, respecting enzymatic catalysis, the reactive sites and neighboring atoms play a more important role than the rest of the molecule when assessing molecular reactivity [7]. Indeed, reactive site-centric information might allow to identify: (1) the metabolic fate and neighbors of a small molecule [8], including metabolic precursors or prodrugs and products of metabolic degradation, (2) small molecules sharing reactivity [9], and (3) competitively inhibited enzymes [10]. Furthermore, neither ML nor mechanistic-based approaches consider the metabolism of the patient, even though the metabolic fate of the drug and the existence of additional targets in the cell might give rise to toxicity. To our knowledge, no available method accounts for human biochemistry when refining the search for drugs.

6.1.3 NICEdrug.ch

Here, we present the development of the NICEdrug.ch database using a more holistic and updated approach to a traditional mechanistic-based screen by (1) adding a more detailed analysis of drug molecular structures and target enzymes based on structural aspects of enzymatic catalysis and (2) accounting for drug metabolism in the context of human biochemistry. NICEdrug.ch assesses the similarity of the reactivity between a drug candidate and a native substrate of an enzyme based on their common reactive sites and neighboring atoms (i.e., the NICEdrug score) in an analogous fashion as the computational tool BridgIT [7]. It also identifies all biochemical transformations in the cellular metabolism that can modify and degrade a drug candidate using a previously developed reaction prediction tool, termed BIochemical Network Integrated Computational Explorer (BNICE.ch) [11], [12] and the ATLAS of Biochemistry [13], [14]. With NICEdrug.ch, we automatically analyzed the functional, reactive, and physicochemical properties of around 250,000 small molecules to suggest the action mechanism, metabolic fate, toxicity, and possibility of drug repurposing for each compound. We apply NICEdrug.ch to study drug action mechanisms and identify drugs for repurposing related to four diseases: cancer, high cholesterol, malaria, and COVID-19. We also sought for molecules in food, as available in fooDB the largest database of food constituents [15], with putative anti SARS-CoV-2 activity. Finally, we provide NICEdrug.ch as an online resource (https://lcsb-databases.epfl.ch/pathways/Nicedrug/). Overall, NICEdrug.ch combines knowledge of molecular structures, enzymatic reaction mechanisms (as included in BNICE.ch [11], [12], [16]–[19]), and cellular biochemistry (currently human,

Plasmodium, and *Escherichia coli* metabolism) to provide a promising and innovative resource to accelerate the discovery and design of novel drugs.

6.2 Materials and Method

6.2.1 NICEdrug pipeline

Here we briefly explain the pipeline to construct and use the NICEdrug (

Figure 6.1: NICEdrug.ch (1) curates available information and calculates the properties of an input compound; (2) identifies the reactive sites of that compound; (3) explores the hypothetical metabolism of the compound in a cell; (4) stores all functional, reactive, bio-, and physico-chemical properties in open-source database; and (5) allows generation of reports to evaluate (5a) reactivity of a small molecule, (5b) drug repurposing, and (5c) druggability of an enzymatic target.

). The details for each step of pipeline are explained after in a separate section.
To build the initial NICEdrug.ch database, we gathered over 70,000 existing small molecules presumed suitable for treating human diseases from three source databases: KEGG, ChEMBL, and DrugBank (Appendix, Figure 8.3). We eliminated duplicate molecules, curated available information, computed thermodynamic properties, and applied the Lipinski rules [20] to keep only the molecules that have drug-like properties in NICEdrug.ch (Figure 6.1, section 6.2.2). NICEdrug.ch currently includes 48,544 unique small molecules from the source databases.

To evaluate the reactivity of the 48,544 drugs, we searched for all possible reactive sites on each drug with BNICE.ch [12] (Figure 6.1, section 6.2.3). All of the 48,544 drugs contain at least one reactive site and hence might be reactive in a cell. In total, we identified more than 5 million potential reactive sites (183k unique) on the 48,544 molecules and matched them to a corresponding enzyme by assigning them to an Enzyme Commission (E.C.) number. All of these enzymes belong to the human metabolic network. Interestingly, 10.4% of identified reactive sites correspond to the p450 class of enzymes, which are responsible for breaking down compounds in the human body by introducing reactive groups on those compounds, also known as phase I of drug metabolism (Appendix, Figure 8.4 pannel A). The sites that were identified varied greatly from simple and small (i.e., comprising a minimum number of one atom) to more complex sites that covered a large part of the molecule. The biggest reactive site includes 30 atoms (Appendix, Figure 8.4 pannel B).

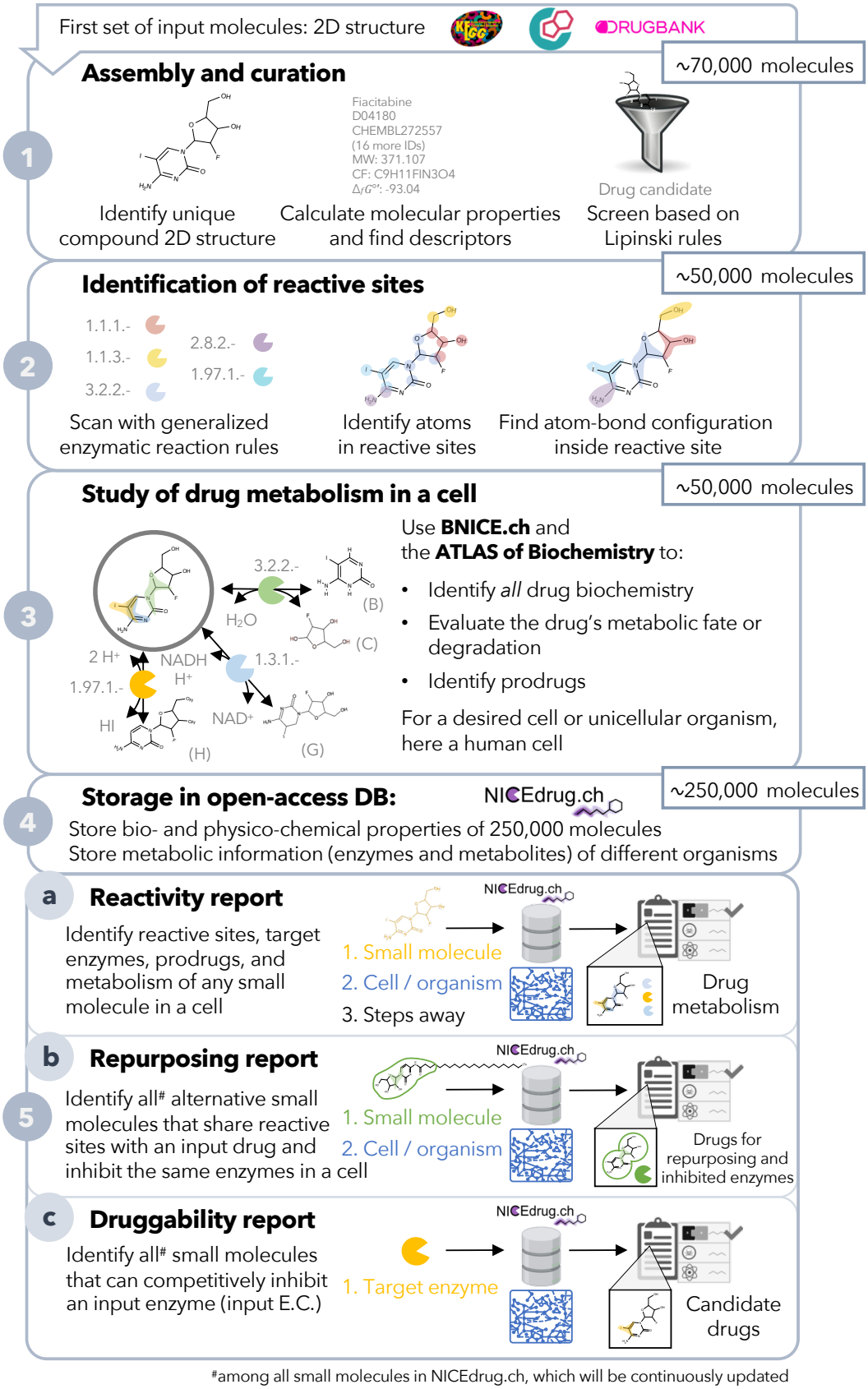


Figure 6.1: NICEdrug.ch (1) curates available information and calculates the properties of an input compound; (2) identifies the reactive sites of that compound; (3) explores the hypothetical metabolism of the compound in a cell; (4) stores all functional, reactive,

bio-, and physico-chemical properties in open-source database; and (5) allows generation of reports to evaluate (5a) reactivity of a small molecule, (5b) drug repurposing, and (5c) druggability of an enzymatic target.

Given the important role of metabolism in the biochemical transformations and toxicity of drugs, we investigated the metabolism of the 48,544 input molecules in human cells. We predicted the hypothetical biochemical neighborhoods of all NICEdrug.ch small molecules in a human cell (i.e., reacting with known human metabolites and cofactors) using a retro-biosynthetic analysis with BNICE.ch (Figure 6.1, section 6.2.4). With this approach, we discovered 197,246 unique compounds connected to the input drugs via one step or reaction (products of the first generation), and the associated hypothetical biochemical neighborhood consists of 630,449 reactions (Appendix, Figure 8.4). The 197,246 unique compounds are part of a new set of bioactive molecules in NICEdrug.ch that might act as drugs or prodrugs in a human cell. We stored the total number of 245,790 small molecules (including the curated set of 48,544 drugs and the new set of 197,246 bioactive compounds), their calculated properties, and biochemistry in our open-access database of drug metabolism, NICEdrug.ch.

To use NICEdrug.ch to identify drug-drug or drug-metabolite pairs that have shared reactivity and target enzymes, we developed a new metric called the NICEdrug score. The NICEdrug score uses information about the structure of the reactive site and its surroundings (as computed using the BridgIT methodology) and is stored in the form of a fingerprint (section 6.2.6). The fingerprint of a molecule's reactive site and the neighborhood around this reactive site—termed the reactive site-centric fingerprint—serves to compare this site-specific similarity with other molecules. We recently showed that the reactive site-centric fingerprint of a reaction provides a better predictive measure of similar reactivity than the overall molecular structure, as the overall structure can be much larger than the reactive site and skew the results by indicating high similarities when the reactivity is actually quite different [7]. Here, we generated reactive site-centric fingerprints for all 20 million reactive sites identified in the 48,544 drugs and 197,246 one-step-away molecules included in NICEdrug.ch. The 20 million reactive site-centric fingerprints for the total 245,790 small molecules are available in NICEdrug.ch to be used in similarity comparisons and classifying molecules (section 6.2.7).

We propose the usage of NICEdrug.ch to generate reports that define the hypothetical reactivity of a molecule, the molecule's reactive sites as identified by target enzymes, and the NICEdrug score between drug-drug and drug-metabolite pairs. The NICEdrug.ch reports can be used for three main applications: (1) to identify the metabolism of small molecules; (2) to suggest drug repurposing; and (3) to evaluate the druggability of an enzyme in a desired cell or organism (Figure 6.1), as we show in the next sections. Currently, NICEdrug.ch includes metabolic information for human cells, a malaria parasite, and Escherichia coli, and it is easily extendible to other organisms in the future.

6.2.2 Curation of input molecules used in the construction of NICEdrug.ch

We constructed the NICEdrug.ch database to gather small molecules suitable for treatment of human diseases. We collected the SMILES structure, synonyms, and any available bio- and physico-chemical property included from three source databases: KEGG, ChEMBL, and DrugBank, which added up to 70,976 molecules by January 2018 (Appenddix, Figure 8.3, panel A). Only molecules that were fully structured were imported to our database. We further curated the imported molecules by removing duplicate structures and merging annotations from different databases into one molecule entry in the database. For removing duplicate structures we used canonical SMIELS [21] generated by openbabel [22] version 2.4.0. This unification method is based on atoms and their connectivity in a molecule in terms of a molecular graph that is captured by the canonical SMILES. Therefore, different resonance forms, stereoisomers, as well as dissociated and charged states of the same compound are mapped to one entry in database. Furthermore, we filtered molecules based on Lipinski rules [20]: (1) the molecular weight should be less than 500 Dalton, (2) the number of hydrogen bond donors should be less than five, (3) the number of hydrogen bond acceptors should be less than ten, and (4) an octanol-water partition coefficient (log P) should be less than five. According to Lipinski rules an active orally drug does not violate more than one of the above criteria. We calculated criteria one, two and three based on the structural information from SMILES of molecules. To assess criterion four, we relied on reported data in the source database. We kept in the NICEdrug.ch database those compounds for which the partition coefficient was not available.

We performed a separate analysis to account for non-unique graph representations of aromatic rings, also called *kekulé structures*. The existence of aromatic rings and the fact that bond-electrons are shared within the ring make several single-double bond assignments possible, which results in multiple kekulé representations for a single molecule (Appendix, Figure 8.3, panel B). We included all such kekulé structures to account for alternative atom-bond connectivity and associated reactivity. We call “effective forms” to the kekulé representations that show different reactive sites than their canonical structures. For example, there can be two effective forms plus the canonical structure (Appendix, Figure 8.3 panel B). In total, we found 42,092 effective forms for 29,994 aromatic compounds in NICEdrug.ch database and we kept them for further analysis.

We also computed the thermodynamic properties of all drugs in NICEdruch.ch. Specifically, we computed the Gibbs free energy of formation ($\Delta_f G'^{\circ}$) using the group contribution method of Mavrovouniotis [23].

The NICEdrug.ch database includes a total number of 48,544 unique and curated small molecules (appendix, Figure 8.3, panel A).

6.2.3 Identification of reactive sites in drugs

The 3D structures of enzyme pockets are complex and mostly unknown. Therefore, evaluating and comparing docking of two small molecules in the pocket of a specific target is impossible most of the times. Using BNICE.ch, we focused on the complementary structure of active sites on substrates, also called *reactive site*. To recognize the potential reactive sites on molecules, we scanned molecules using expert-curated generalized reaction rules of BNCIE.ch [13], which mimic the identification of substrates by the enzyme pocket and account for the promiscuous activity of enzymes. Theses reaction rules incorporate the information of biochemical reactions and have third-level Enzyme Commission (EC) identifiers. Each BNICE.ch reaction rule accounts for three levels of information: (1) atoms in reactive sites of compounds, (2) connectivity and configuration of atom bonds in the reactive site, and (3) mechanism of bond breakage and formation during the reaction. As of May 2020, BNICE.ch contains 450 bidirectional generalized reaction rules that can reconstruct 8118 KEGG reactions [13]. Here, we include all BNICE.ch rules to identify all possible reactive sites on a given drug in two steps. First, a BNICE.ch rule identifies all atoms in a compound that belong to the rule's reactive site. Second, the rule evaluates the connectivity of the atoms previously identified. The candidate compounds for which a BNICE.ch rule identified a reactive site were validated as metabolically reactive and considered for analysis in NICEdrug.ch.

It is important to note that thanks to the generalized reaction rules, which abstract the knowledge of thousands of biochemical reactions, BNICE.ch is able to reconstruct known biotransformations and also propose novel metabolic reactions. This was demonstrated in the reconstruction of the ATLAS of Biochemistry [13], which involves up to 130,000 reactions between known compounds.

6.2.4 Analysis of drug metabolism in human cells.

To mimic biochemistry of human cells and simulate human drug metabolism, we collected all available information (metabolites and metabolic activities or EC numbers of enzymes) on human metabolism from three available databases: the human metabolic models Recon3D [24] and HMR [25], and the Reactome database [26]. These three databases include a total of 2,266 unique human metabolites and 2,066 unique EC numbers of enzymes (appendix, Table 8.22).

To explore the biochemical space beyond the known human metabolic reactions and compounds, we used (1) the generalized enzymatic reaction rules of BNICE.ch that match up to the third EC level the collected human enzymes, and (2) all of the collected human metabolome. We evaluated the reactivity of each drug in a human cell using the retro-biosynthesis algorithm of BNICE.ch, which predicts hypothetical biochemical transformations or *metabolic neighborhood* around the drug of study. We generated with BNICE.ch metabolic reactions in which each drug and all known human metabolites could participate as substrate or

products. We also allowed a set of 53 known cofactors to react with the human metabolites (appendix, Table 8.22).

We define the boundaries of the metabolic neighborhood of a molecule with a maximum number of reactions or *steps away* that separate the input molecule (drug of study) from the furthest compound. In BNICE.ch, a generation *n* of compounds involves all metabolites that appear for the first time in the metabolic neighborhood of a drug after *n* reactions or steps happened. For example, in the case study of 5-FU we find the compound 5-Fluorouridine in generation 2 or 2 steps away, which means there are two metabolic reactions that separate 5-FU and 5-Fluorouridine (section 6.3.1).

In NICEdrug.ch, there exist 197,246 compounds in generation 1 (1 step away) from all input drugs. The 197,246 compounds are part of the potential drug metabolic neighborhood in human cells. Out of all generation 1 molecules, 13,408 metabolites can be found in human metabolic models and HMDB database [27], 16,563 metabolites exist in other biological databases, and the remaining 167,245 metabolites are catalogued as known compounds in chemical databases (i.e., PubChem). Note that HMDB includes native human metabolites and non-native human compounds, like food ingredients.

The 197,246 products that are one-step away of all NICEdrug.ch molecules are part of a hypothetical biochemical neighborhood of 630,449 drug metabolic reactions. Of all drug metabolic reactions, 5,306 reactions are cataloged in biological databases, and the remaining 625,143 reactions are novel. A majority of the reactions involved oxidoreductases (42.54%), broken down into 27.45% of lyases, 7.15% of hydrolases, 6.28% of transferases, 1% of isomerases, and 15.58% of ligases. Interestingly based on the previously identified reactive sites, out of the 265,935 (42.54% of 625,143) oxidoreductase reactions, 49.92% are catalyzed by the p450 family of enzymes, which are known to be responsible for the metabolism of drug (Appendix, Figure 8.4 panel C).

6.2.5 Using NICEdrug.ch database for analysis of the metabolic neighborhood of a drug

In NICEdrug.ch webserver, users can look up for a drug using the drugs' name and other identifiers like ChEMBL, DrugBank and KEGG. NICEdrug.ch will report a unique identifier for the compound that will be input for upcoming analysis modules. The *reactivity* report allows to study the metabolic network around an input molecule. The input to this module is: (1) the unique identifier of the drug of interest, (2) a maximum number of reactions or *steps away* that shall separate the input drug to the furthest compound in the metabolic neighborhood.

The output of this analysis is a report in the form of a tsv file that includes all compounds and metabolic reactions in the metabolic neighborhood of the input drug. One can also export the neighborhood in the form of a visual graph, in which nodes are molecules and edges are reactions.

6.2.6 Definition of the NICEdrug score

Based on the theory of lock and key, two metabolites that can be catalyzed by the same enzyme may have similar reactive sites and also neighboring atoms. In order to quantify the similarity inside and around reactive sites of two molecules, we developed a metric called *NICEdrug score* (Figure 6.2), which is inspired on BridgIT [7]. BridgIT assesses the similarity of two reactions, considering the reactive site of the participating substrates and their surrounding structure until the seventh atom out of the reactive site.

The NICEdrug score is an average of two similarity evaluations: (1) the atom-bond configuration inside reactive site (α parameter), and (2) the 7 atom-bond chain molecular structure around the reactive site (β parameter). The NICEdrug score, and its parameters α and β, range between 0 and 1 when they indicate no similarity and identical structure, respectively. Different constraints on the α and β parameters determine the identification of different types of inhibition like para-metabolites and anti-metabolites (see section 6.2.8 and 6.2.9).

We show the evaluation of NICEdrug scores for three example compounds (Figure 6.2). In this example, Digoxin, Labriformidin and Lanatoside C all share the reactive site corresponding to EC number 5.3.3.- (α=1). Starting from the atoms of the identified reactive site, eight description layers of the molecule were formed, where each layer contains a set of connected atom-bond chains. Layer zero includes types of atoms of reactive site and their count. Layer 1 expands one bond away from all of the atoms of reactive site and accounts for atom-bond-atom connections. This procedure is continued until layer 7, which includes the sequence of 8 atoms connected by 7 bonds. Then, we compare the fingerprint of each molecule to the other participants of the class based on the Tanimoto similarity scores. A Tanimoto score near 0 designates no or low similarity, whereas a score near 1 designates high similarity in and around reactive site. Lanatoside C and Digoxin share the same substructure till 8 layers out of reactive site which is presented in the NICEdrug score by preserving score 1 in all layers, so the overall Tanimoto score for these two compounds in the context of EC number 5.3.3.- is 1 (α=1 and β=1). However, the structure of two compounds are not exactly the same and actually Lanatoside C has 8 more carbon atoms and 6 more oxygen atoms, shaped as an extra benzenehexol ring and an ester group. Although this part is far from the reactive site, based on the NICEdrug score they both can perfectly fit inside the binding pocket of a common protein related to this reactive site. This hypothesis is proved by experiments reported in KEGG and DrugBank. According to DrugBank and KEGG, Lanatoside C has actions similar to Dioxin and both of them have the same target pathways: Cardiac muscle

contraction and Adrenergic signaling in cardiomyocytes. Furthermore, target protein for both of them is ATP1A.

Also, the NICEdrug score effectively captures and quantifies differences around the reactive site. The substructure around the reactive site in Labriformidin is slightly different (α=1 and β <1). The difference is calculated trough different layers of the NICEdrug score.

In the case study of 5-FU (see section 6.3.1), in order to predict competitive inhibition, we analyzed all the metabolites that share reactive site with 5-FU or its downstream products (α=1) and then we ranked the most similar metabolites based on their similarity in neighborhood of reactive site to 5-FU or its downstream products (β). To assess the structural differences in the reactive sites themselves (α), we implemented the Levenshtein edit distance algorithm [28] to determine how many deletions, insertions, or substitutions of atom/bonds are required to transform one pattern of reactive site into the other one. Here, the edit distance explains the difference between the reactive sites of the intermediate and the human metabolite. However, even slight changes in the reactive site affect its interaction with the binding site. To ensure that the divergence retained the appropriate topology, we compared the required edit on reactive site with interchangeable groups, termed bioisosteric groups [29]. These bioisosteric groups contain similar physical or chemical properties to the original group and largely maintain the biological activity of the original molecule. An example of this is the replacement of a hydrogen atom with fluorine, which is a similar size that does not affect the overall topology of the molecule. For this analysis, we used 12 bioisosteric groups adapted from the study by Papadatos et.al. [29].

To predict irreversible Inhibitors in metabolism of 5-FU, we kept only molecules with a similarity score greater than 0.9 to metabolites (β>0.9), to preserve a high similarity in the neighborhood of the reactive sites. Then, we checked which ones contained reactive sites that differed only in the replacement of bioisosteric groups (α~1).

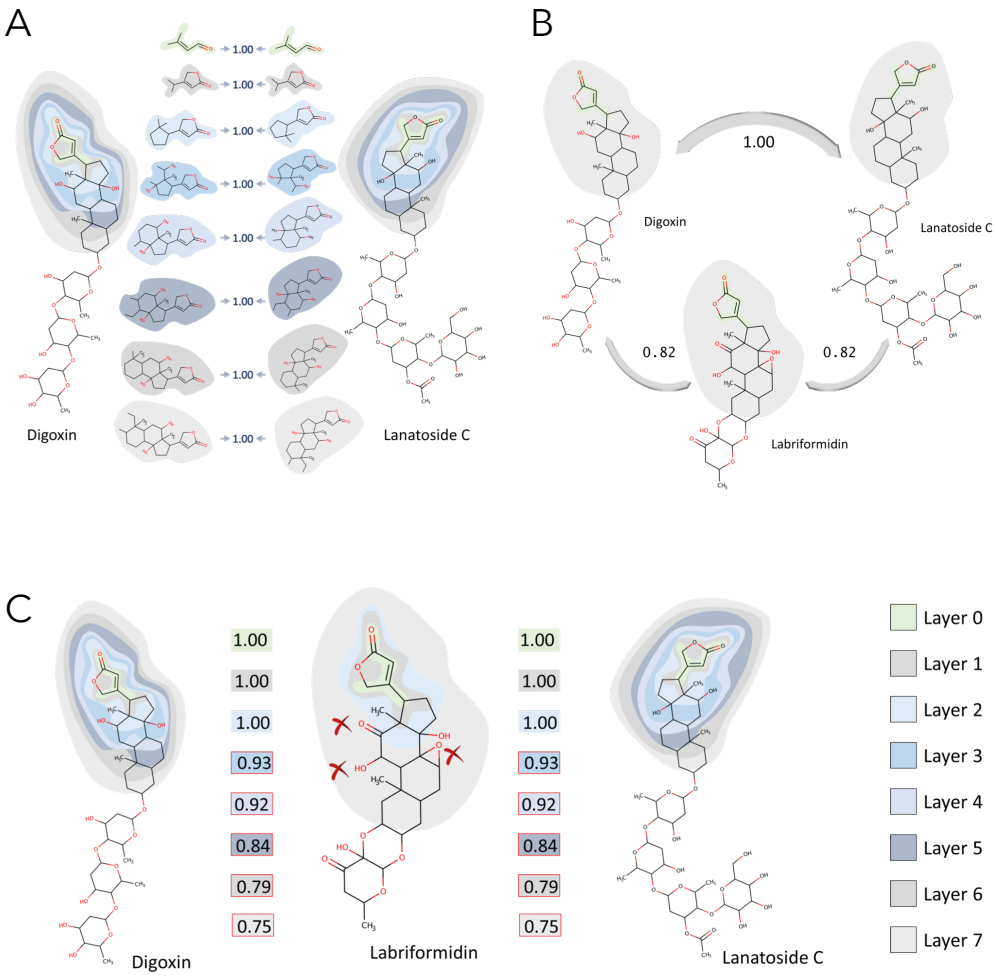


Figure 6.2: Example of NICEdrug score calculation. The NICEdrug score takes into account the structure of a molecule's reactive site and its seven-atom-away neighborhood for similarity evaluation, analogous to BridgIT.

6.2.7 Classification of drugs based on the NICEdrug score

Classification of compounds with similar structure is normally used to assign unknown properties to new compounds. For instance, one can infer ligand-protein binding for a drug when its action mechanism or the structure of the target proteins are not known. In this study, we have considered four strategies to classify drugs, which are from less to more stringent: classifying (1) molecules that participate in reactions with the same EC up to the 3th level, (2) molecules that in addition share a BNICE.ch reaction rule, (3) molecules that in addition to both previous points share reactive site, (4) molecules that show high similarity of reactive site and neighborhood based on the NICEdrug score.

The EC number guarantees that molecules are catalyzed with similar overall reaction mechanism. Generalized reaction rules from BNICE.ch further capture different submechanisms inside an EC number [13]. A BNICE.ch reaction rule might involve more than one reactive site. Hence, information of reactive sites

provides further insights into the molecule's reactivity. Furthermore, similarity of reactive sites and their neighborhoods based on the NICEdrug score increase the comparison resolution and this is the basis of the classification in NICEdrug.ch.

In NICEdrug.ch database there exist 95,342 classes that comprise all drugs and human compounds sharing EC, BNICE.ch rule, and reactive site (classification based on our strategy 3). We computed the NICEdrug score between all pairs of molecules in a class and this information is available in NICEdrug.ch.

6.2.8 Identification of drugs acting as para-metabolites based on NICEdrug score

Small molecules that share reactive site and are structurally similar to native human metabolites enter and bind the pocket of native enzymes and competitively inhibiting catalysis acting as para-metabolites [30]. In this study, we define as para-metabolite any drug or any of its metabolic neighbors that (1) shares reactive site with native metabolites (α=1), and (2) preserves a high NICEdrug score with respect to the reactive site neighborhood (β>0.9).

6.2.9 Identification of drugs acting as anti-metabolites based on NICEdrug score

Small molecules that do not share reactive site but are structurally similar to native human metabolites might enter the binding pocket of native enzymes and inhibiting catalysis acting as anti-metabolites [30]. In this study, we define as anti-metabolite any drug or any of its metabolic neighbors that (1) differs slightly in reactive site from a native metabolite (α~1), and (2) preserves high similarity in the reactive site neighborhood (β>0.9). We hypothesize that a low divergence in the reactive site, still allows a non-native compound to enter and bind the enzyme pocket since it is structurally similar enough to the native substrate.

6.2.10 Identification of NICEdrug toxic alerts

We obtained all NICEdrug toxic alters from ToxAlert database [31]. ToxAlert database includes about 1,200 structural toxic alerts associated with particular types of toxicity. Toxic alerts are provided in the form of SMART patterns that are searchable in SMILES structure of molecules. NICEdrug.ch uses openbable tool [22] to search for these structural alerts on SMILES of compounds.

6.2.11 Collection of reference toxic molecules in NICEdrug.ch

Studying the adverse effects of chemicals on biological systems has led to development of databases cataloging toxic molecules. The Liver Toxicity Knowledge Base (LTKB) integrates 1,036 molecules annotated with human Drug-induced liver injury risk (severity). Super toxic DB include about 60k toxic molecules, that

are annotated with their toxicity estimate, LC_{50}/LD_{50} i.e., lethal dose or concentration at which 50% of a population dies.

As a resource of approved toxic molecules, we collected all of the molecules cataloged as toxic in LTKB and super toxic databases. We used this collection as a reference to compare the similarity of drugs or and products of drug metabolism with approved toxic molecules.

6.2.12 Definition of a toxicity score in NICEdrug.ch

The number of molecules labeled as toxic in databases is disproportionally low compared to the space of compounds. On the other hand, toxic alerts are defined for a big number of compounds and are linked to redundant molecular structures.

We measured the similarity of drugs and their metabolic neighbors with the collection of reference toxic molecules using the NICEdrug score. We assigned toxic alerts to molecules in NICEdrug.ch if a molecule and toxic molecule shared a molecular substructure linked to the toxic alert.

Finally, NICEdrug.ch provides a toxicity report in the form of a csv file for each molecule in the metabolic neighborhood including six values linked to the most similar toxic molecules in both toxic reference databases (LTKB and supertoxic databases): (1) the NICEdrug score between the drug and those most similar toxic molecules, (2) the severity degree of the hepatotoxic compound, and log(LC_{50}) of the supertoxic compound, and (3) the number of common toxic alerts between the drug and the most similar toxic molecules. The list of toxic alerts is also provided.

We combined the six values of the toxicity report into a toxicity score defined as follows:

$$\sum_{i} \text{NICEdrug score} \times (\log(\text{LC}_{50}) \text{ or severity degree}) \times \text{number of common NICEdrug toxic alerts}$$

$$\text{i} \in \{\text{the most similar approved toxic molecules in LTKB and supertoxic databases}\}$$

The toxicity score in NICEdrug.ch served to quantify the toxicity of each molecule in the metabolic neighborhood of a drug, recapitulate known toxic molecules, and suggest new toxic compounds.

6.2.13 Analysis of essential enzymes and linked metabolites in *Plasmodium* and human cells

We extracted information of essential genes and enzymes for liver-stage malaria development from our recent study [32]. In this study, we developed the genome-scale metabolic model of *Plasmodium berghei*, which shows high consistency (approximately 80%) with the largest gene knockout datasets in *Plasmodium*

blood [33] and liver stages [32]. There are 178 essential genes for *P. berghei*'s growth simulating liver-stage conditions [32]. Here, we identified the substrates of those essential metabolic enzymes, which comprise a set of 328 metabolites (appendix, Table 8.26). To further minimize on the host cell, we filtered out those *Plasmodium* enzymes that share 4th level E.C. with human essential enzymes. We used available CRISPR gene essentiality data in various human cell lines [34] to identify essential genes and enzymes in human cells (appendix, Table 8.26). We further identified essential metabolites in human cells (appendix, Table 8.26) using the latest human genome-scale metabolic model [35] and the metabolic information associated to the essential human genes. Subtracting essential parasite and human enzymes resulted in the analysis of 32 essential *Plasmodium* enzymes catalyzing 68 metabolites and 157 unique metabolite-enzyme pairs in the parasite (appendix, Table 8.27).

6.2.14 Identification of drugs to target malaria and minimize side effects on human cells

Those molecules that themselves and their downstream products cannot act as inhibitors of essential metabolic enzymes in the human host cell while they can target essential *Plasmodium* enzymes are attractive antimalarial candidates.

We first used NICEdrug.ch to look for small molecules that share reactive site with the 32 essential *Plasmodium* enzymes and they have good similarity score in reactive site neighborhood to native substrates of essential enzymes of parasite, i.e. NICEdrug score above 0.5 (appendix, Table 8.27). We also identified prodrugs that might lead to downstream products with similar reactive site and neighborhood (NICEdrug above 0.5) to any of the essential *Plasmodium* metabolites (appendix, Table 8.27). We suggest those drugs and downstream products act as antimetabolites and competitively inhibit the essential enzymes in the parasite. Overall, we identified 516 drugs that directly compete with essential metabolites and 1,164 prodrugs that need to be biochemically modified between one to three times in human cell to render inhibition of essential enzymes.

We next combined information of essential *Plasmodium* and human metabolites to screen further the drug search using NICEdrug.ch. Out of the hypothetical 516 antimalarial candidates, we identified 64 drugs that share reactive site with parasite metabolites (NICEdrug score above 0.5) and not with human metabolites (NICEdrug score below 0.5), making them good candidates for drug design (appendix, Table 8.27).

6.2.15 Prediction of inhibitors among food based molecules

We used the reactive site-centric fingerprint available in NICEdrug.ch to identify molecules in food that share reactive site with native substrates of human enzymes and hence might inhibit those enzymes. We retrieved the total set of 80,000 compounds from FooDB [15], and treated them as input molecules into the NICEdrug

pipeline (Figure 6.1) to identify reactive sites and evaluate their biochemistry, as done for all molecules in NICEdrug.ch.

6.2.16 Identification of small molecules to target COVID-19

A recent study reported 332 host factors of SARS-CoV-2 [36]. Out of the 332 proteins, 97 have catalytic function and EC number assigned, and are potential targets of small molecules. We evaluated the druggability of these 97 enzymes using NICEdrug.ch.

To generate a druggability report, NICEdrug.ch first gathers the metabolic reactions associated with the protein EC numbers. NICEdrug.ch uses 11 databases (including HMR, MetaCyc, KEGG, MetaNetX, Reactome, Rhea, Model SEED, BKMS, BiGG models and Brenda) as source of metabolic reactions. All these databases involve a total of 60k unique metabolic reactions.

Out of the 97 host factor enzymes, we identified 22 enzymes that are linked to fully-defined metabolic reactions. Fully-defined metabolic reactions fulfill three criteria. (1) There is a secondary structure available for all the reaction participants, which means there are available mol files. (2) There is a fully defined molecular structure for all the reaction participants, which means molecules with unspecified R chains are discarded. (3) There is a BNICE.ch enzymatic reaction rule assigned to the reaction (appendix, Table 8.28).

NICEdrug.ch identified 22 host factor enzymes with 24 unique linked EC numbers and 145 unique fully defined reactions. NICEdrug.ch extracts the metabolites participating in these reactions and identifies their reactive site for a reactive-site centric similarity evaluation against a list of molecules. To this end, NICEdrug.ch reports the list of molecules ranked based on the NICEdrug score. The molecule with the highest NICEdrug score shares the highest reactive site-centric similarity with the native substrate of the target enzyme (appendix, Table 8.28).

We found 1,301 molecules that show NICEdrug score above 0.5 with respect to substrates of the 22 SARS-CoV-2 hijacked enzymes (appendix, Table 8.28). Out of 1,301 molecules, 465 are drugs cataloged in DrugBank, KEGG drugs or ChEMBL databases, 712 are active molecules one step away of 1,419 prodrugs, and 402 are food molecules (appendix, Table 8.28).

To better understand the classes of drugs or food molecules, we classified drugs based on their KEGG drug groups (Dgroups) and food molecules based on their food source. Out of 465 drugs identified, 43 drugs are assigned to 55 different Dgroups and 402 food molecules belong to 74 different food sources (appendix, Table 8.28).

6.3 Results and Discussion

6.3.1 NICEdrug.ch suggests inhibitory mechanisms of the anticancer drug 5-FU and avenues to alleviate its toxicity.

As a case study, we used NICEdrug.ch to investigate the mode of action and metabolic fate of one of the most commonly used drugs to treat cancer, 5-fluorouracil (5-FU), by exploring its reactivity and the downstream products or intermediates that are formed during the cascade of biochemical transformations. 5-FU interferes with DNA synthesis as an antimetabolite [37], meaning that its various intermediates like 5-fluorodeoxyuridine monophosphate (FdUMP) are similar enough to naturally occurring substrates and they can act as competitive inhibitors in the cell.

We therefore used NICEdrug.ch to study the intermediates of 5-FU that occurred between one to four reaction steps away from 5-FU (appendix, Table 8.23), which is a reasonable range to occur in the body after 5-FU treatment [38]. This analysis identified 407 compounds (90 biochemical and 317 chemical molecules) that have the biochemical potential to inhibit certain enzymes. Because the NICEdrug score that analyses reactive site and neighborhood similarities can serve as a better predictor of metabolite similarity, we assessed the NICEdrug score of the intermediates compared to human metabolites. This resulted in a wide range of NICEdrug scores between the different 5-FU intermediates and human metabolites, ranging from no similarity at a NICEdrug score of 0 to the equivalent substructure on a compound at a NICEdrug score of 1. More importantly, some of the 407 metabolite inhibitors (as explained next) were known compounds that have been investigated for their effects on 5-FU toxicity, but most of these compounds were newly identified by NICEdrug.ch and could therefore serve as avenues for future research into alleviating the side effects of this drug.

We investigated these 407 compounds in more detail, looking first at the set of already validated metabolite inhibitors. 5-Fluorouridine (two steps away from 5-FU) and UDP-L-arabinofuranose (four steps away from 5-FU) are very similar to uridine, with NICEdrug scores of 0.95 and 1, respectively. Uridine is recognized as a substrate by two human enzymes, cytidine deaminase (EC: 3.5.4.5) and 5'-nucleotidase (EC: 3.1.3.5) (Figure 6.3). Therefore, NICEdrug.ch predictions show that the degradation metabolism of 5-FU generates downstream molecules similar to uridine, which likely leads to the inhibition of these two enzymes. This effect has already been investigated as a potential method for reducing the toxicity of 5-FU, wherein it was proposed that high concentrations of uridine could compete with the toxic 5-FU metabolites [39].

NICEdrug.ch also identified a few potential metabolites that have not been previously studied for their effects. These metabolites share a reactive site with native human metabolites and differ in the reactive site neighborhood, and we refer to them as *para-metabolites* [40]. 6-Methyl-2'-deoxyadenosine, purine-deoxyribonucleoside, and 2'-deoxyisoguanosine structurally resemble the reactive site neighborhood of

deoxyadenosine, with respective NICEdrug scores of 1, 1, and 0.91. Similarly, 2-aminoadenosine, 2-chloroadenosine, and 2-methylaminoadenosine (four steps from 5-FU) have the same reactive site neighborhood as adenosine, with NICEdrug scores of 1, 1, and 0.96, respectively. Adenosine and deoxyadenosine are both native substrates of the adenosine kinase (EC: 2.7.1.20) and 5'-nucleotidase (EC: 3.1.3.5) (Figure 6.3). Therefore, we suggest that the 5-FU derivatives 2-aminoadenosine and 2-chloroadenosine are competitive inhibitors for the two enzymes adenosine kinase and 5'-nucleotidase. With these new insights from NICEdrug.ch, we hypothesize that co-administering adenosine or deoxyadenosine and uridine (Figure 6.3) with 5-FU might be required to reduce its toxic effects and hopefully alleviate the side effects of the 5-FU cancer treatment.

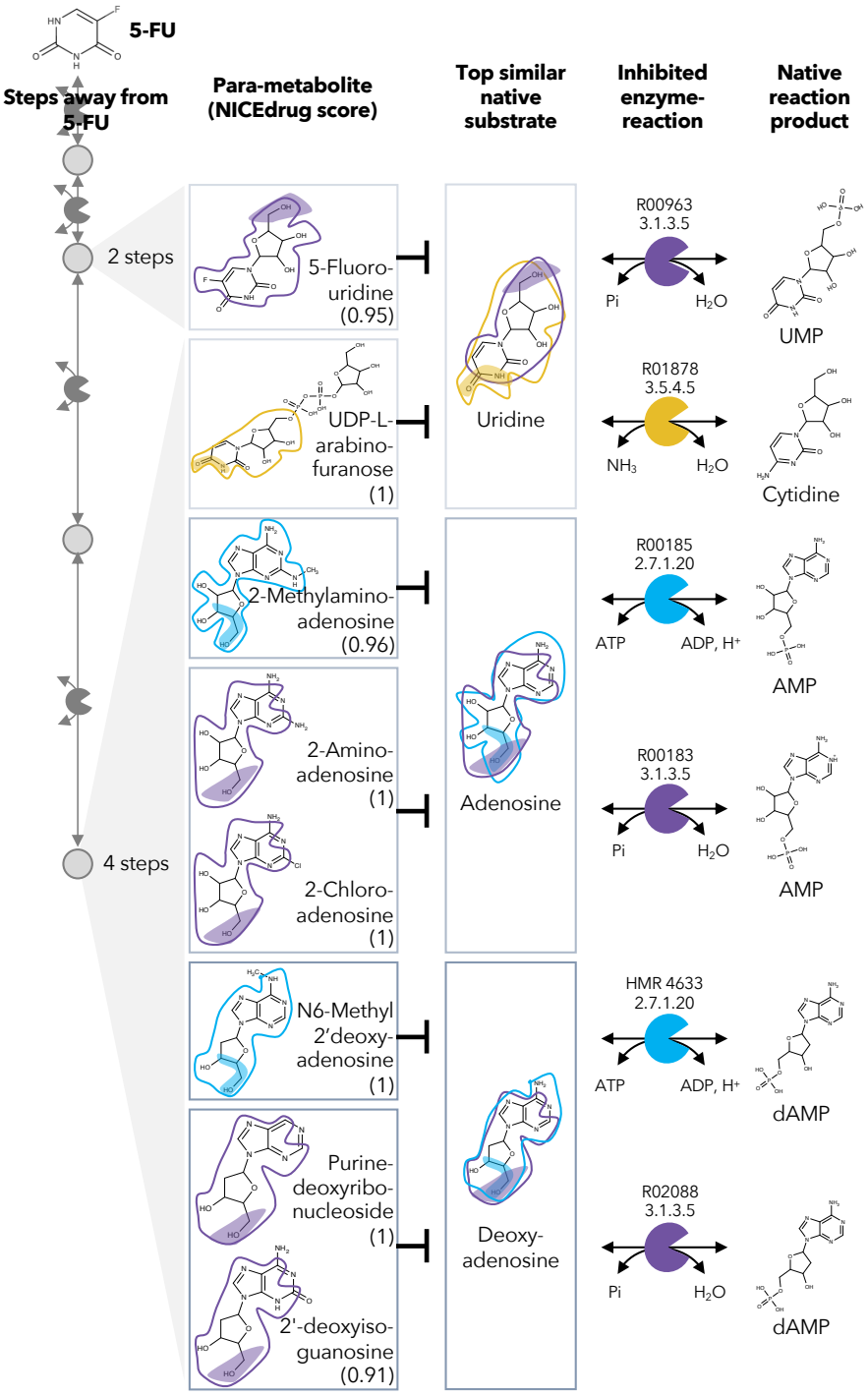


Figure 6.3: Similarity in reactive site and neighborhood defines para-metabolites in 5-FU metabolism and inhibited human metabolic enzymes. Eight para-metabolites in the 5-FU metabolic neighborhood (represented as defined in section 6.2.8). We show the most similar native human metabolites, inhibited enzymes, and native products of the reactions.

6.3.2 Metabolic degradation of 5-FU leads to compounds with Fluor in their reactive site that are less reactive and more toxic than other intermediates

In the previous case study, we showed inhibitors that contain the identical active site to the native enzyme. However, a slightly different reactive site might still be able to bind to an enzyme and compete with a native substrate, also defined as *anti-metabolite* [41]. We explored this scenario by defining relaxed constraints in two steps. We first identified all atoms around a reactive site to compare the binding characteristics between the native molecule and putative inhibitor. Next, we compared the reactive site of the native molecule and putative inhibitor and scored the latter based on similarity (section 6.2.9). Following these two steps, we assessed the similarity between intermediates in the 5-FU metabolic neighborhood and human metabolites. Among all 407 compounds in the 5-FU metabolism (appendix, Table 8.23), we found 8 that show a close similarity to human metabolites (NICEdrug score above 0.9,

Figure 6.4) that might be competitive inhibitors or anti-metabolites. Inside the reactive site, the original hydrogen atom is bioisosterically replaced by fluorine. F-C bonds are extremely stable and therefore block the active site by forming a stable complex with the enzyme. The inhibitory effect of the intermediates tegafur, 5-fluorodeoxyuridine, and F-dUMP (one to two reaction steps away) has been confirmed in studies by Kobayakawa et.al [42] and Bielas et.al [43]. In addition, NICEdrug.ch also predicts that 5flurim, 5-fluorodeoxyuridine triphosphate, 5-fluorodeoxyuridine triphosphate, 5-fluorouridine diphosphate, and 5-fluorouridine triphosphate, some of which occur further downstream in the 5-FU metabolism, also act as antimetabolites (

Figure 6.4). Based on the insights from NICEdrug.ch, we suggest the inhibitory and side effect of 5-FU treatment might be more complex than previously thought. 5-FU downstream products are structurally close to human metabolites and might form stable complexes with native enzymes. This knowledge could serve to further refine the pharmacokinetic and pharmacodynamic models of 5-FU and ultimately the dosage administered during treatment.

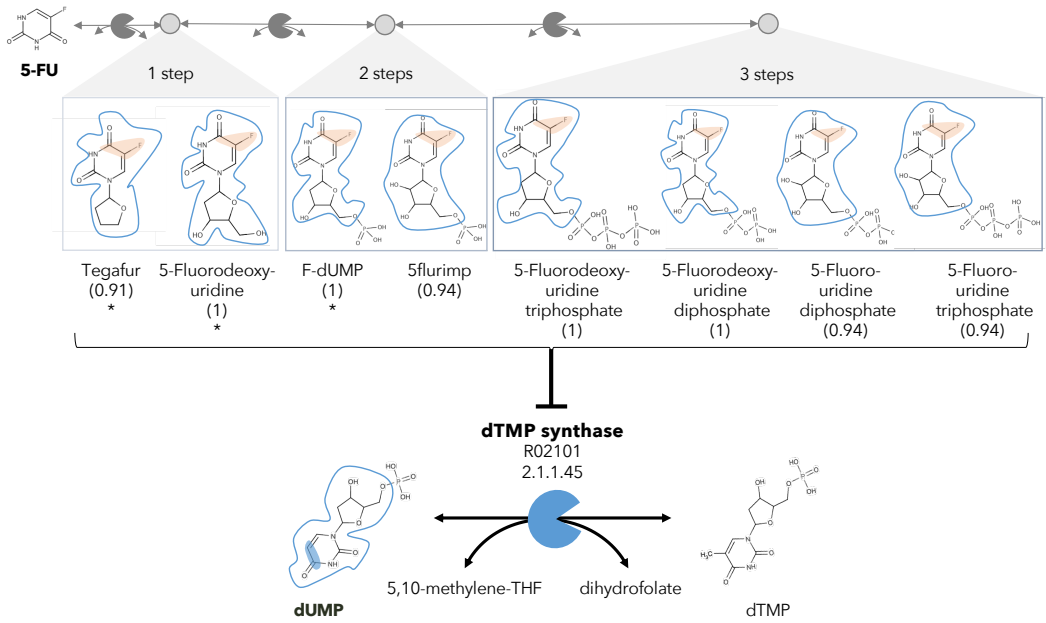


Figure 6.4: A different reactive site but similar neighborhood defines top anti-metabolites in 5-FU metabolism and inhibited human metabolic enzyme. Eight anti-metabolites of dUMP in the 5-FU metabolic neighborhood (represented as defined in 6.2.9). Note that the reactive site of the anti-metabolites is different than the one of the native human metabolite, but the neighborhood is highly similar, which determines the high NICEdrug score (value in parenthesis). We show the inhibited human enzyme (dTMP synthase) and reaction, and its native product.

6.3.3 NICEdrug.ch identifies toxic alerts in the anticancer drug 5-FU and its products from metabolic degradation.

The concept of drug toxicity refers not to overdoses but instead to the toxic effects at medical doses [44], which often occur due to the degradation products generated through drug metabolism. Extensive efforts have been expended to identify toxic molecules or, more generally, to extract the substructures that are responsible for toxicity (called structural alerts). The Liver Toxicity Knowledge Base (LTKB) and the super toxic database include 1,036 and about 60k toxic molecules, respectively [45], [46]. ToxAlert provides around 1,200 alerts related to different forms of toxicity [31]. However, the number of molecules that are analyzed and labeled as toxic in databases is disproportionally low compared to the space of compounds. Additionally, structural alerts are indicated for many compounds, and current alerts might identify redundant and over-specific substructures, which questions their reliability [47].

To quantify the toxicity of downstream products of drugs in NICEdrug.ch, we collected all of the molecules cataloged as toxic in the LTKB and super toxic databases (approved toxic molecules) along with their lethal dose (LC_{50}), as well as the existing structural alerts provided by ToxAlert. We measured the similarity of an

input molecule with all approved toxic molecules using the reactive site-centric fingerprints implemented in BridgIT and the NICEdrug score (see section 6.2.6). Next, we scanned both the toxic reference molecule and the input molecule for structural hints of toxicity, referred to here as *NICEdrug toxic alerts*. We kept common NICEdrug toxic alerts between the reference, which is a confirmed toxic compound, and input molecule. With this procedure in place, NICEdrug.ch finds for each input molecule the most similar toxic molecules along with their common toxic alerts and serves to assess the toxicity of a new molecule based on the mapped toxic alerts. Additionally, the NICEdrug toxic alerts and toxicity level of drug intermediates can be traced with NICEdrug.ch through the whole degradation pathway to reveal the origin of the toxicity.

As an example, we herein tested the ability of NICEdrug.ch to identify the toxicity in 5-FU metabolism. First, we queried the toxicity profile of all intermediates in the 5-FU metabolic neighborhood, integrating both known and hypothetical human reactions (see section 6.2.4). In this analysis, we generated all compounds up to four steps away from 5-FU. Based on the toxicity report of each potential degradation product, we calculated a relative toxicity metric that adds the LC_{50} value, NICEdrug score, and number of common NICEdrug toxic alerts with all approved toxic drugs (see section 6.2.12). We generated the metabolic neighborhood around 5-FU, and labeled each compound with our toxicity metric (appendix, Table 8.23). Interestingly, we show that the top most toxic intermediates match the list of known three toxic intermediates in 5-FU metabolism (Figure 6.5) [48]. Based on the toxicity analysis in NICEdrug.ch for 5-FU, we hypothesize there are highly toxic products of 5-FU drug metabolism that had not been identified either experimentally or computationally and it might be necessary to experimentally evaluate their toxicity to recalibrate the dosage of 5-FU treatment.

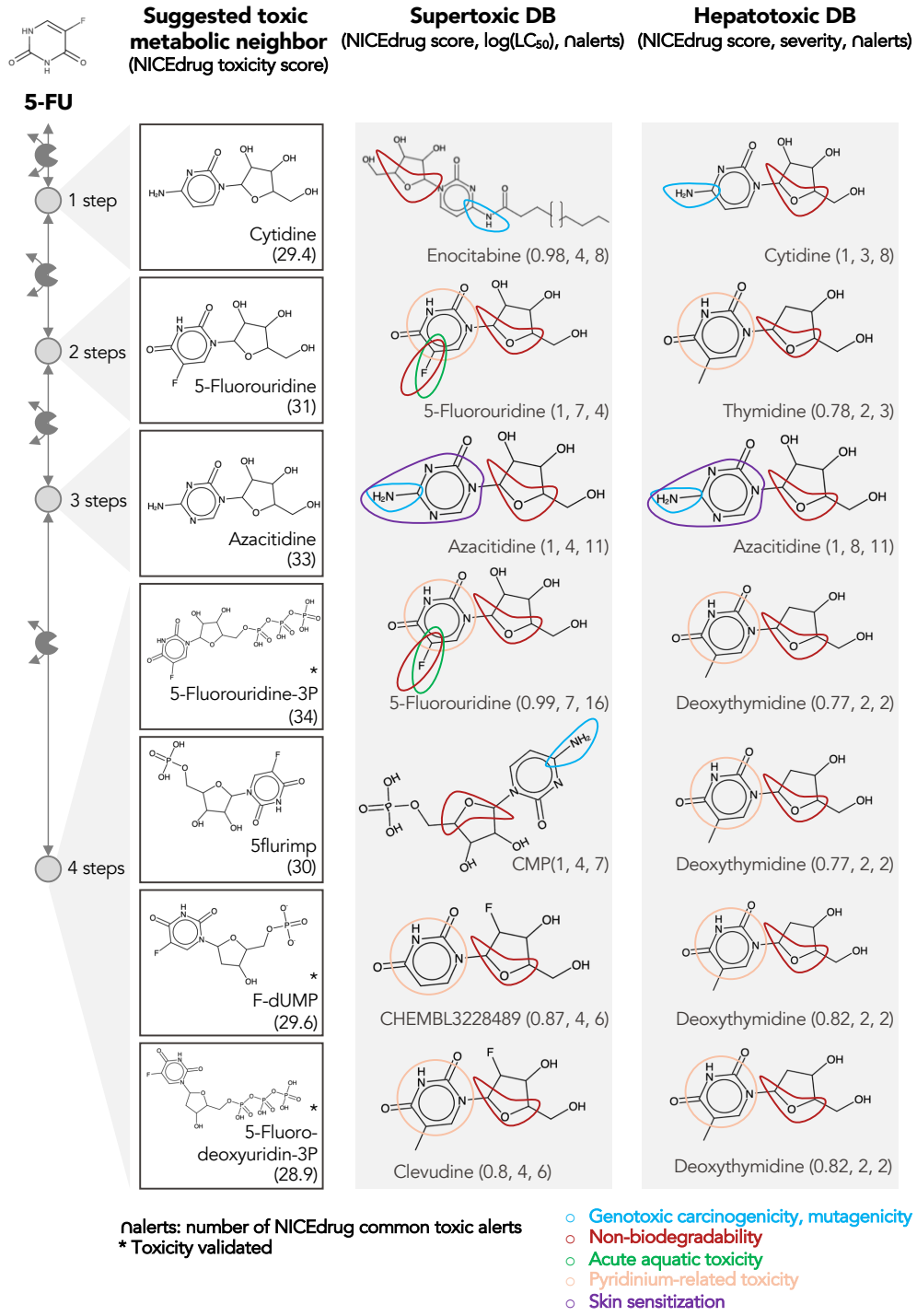


Figure 6.5: Comparing downstream products to known toxic molecules and analyzing their common structural toxic alerts explains metabolic toxicity of 5-FU. Example of six suggested toxic molecules in the 5-FU metabolic neighborhood (represented as defined in 6.2.12). We show toxic compounds from the supertoxic and hepatotoxic databases that lead to the highest NICEdrug toxicity score (number under toxic intermediate name). We highlight functional groups linked to five NICEdrug toxic alerts (legend bottom right).

6.3.4 The NICEdrug reactive site-centric fingerprint accurately clusters statins of type I and II and guides drug repurposing.

Because potential side effects of a drug are documented when the drug passes the approval process, repurposing approved drugs for other diseases can reduce the medical risks and development expenses. For instance, the antitussive noscapine has been repurposed to treat some cancers [49], [50]. Because NICEdrug.ch can search for functional (i.e., reactivity), structural (i.e., size), and physicochemical (i.e., solubility) similarities between molecules while accounting for human biochemistry, we wanted to determine if NICEdrug.ch could therefore suggest drug repurposing strategies.

As a case study, we investigated the possibility of drug repurposing to replace statins, which are a class of drugs often prescribed to lower blood cholesterol levels and to treat cardiovascular disease. Indeed, data from the National Health and Nutrition Examination Survey indicate that nearly half of adults 75 years and older in the United States use prescription cholesterol-lowering statins [51]. Since some patients do not tolerate these drugs and many still do not reach a safe blood cholesterol level [52], there is a need for alternatives. Being competitive inhibitors of the cholesterol biosynthesis enzyme 3-hydroxy-3-methyl-glutaryl-coenzyme A reductase (HMG-CoA reductase) [53], [54], all statins share the same reactive site. BNICE.ch labeled this reactive site, in a linear or circular form, as corresponding to an EC number of 4.2.1.- [55]. NICEdrug.ch includes 254 molecules with the same reactive site that are recognized by enzymes of E.C. class 4.2.1.-, ten of which are known statins. We used the NICEdrug score to cluster the 254 molecules into different classes (appendix, Table 8.24, Figure 6.6). Two of the classes correspond to all currently known statins, which are classified based on their activity into type 1 and 2, wherein statins of type 2 are less active and their reactive site is more stable compared to type 1. This property is well distinguished in the clustering based on the NICEdrug score (Figure 6.6 panel A).

In addition to properly classifying the ten known statins (Figure 6.6 panel B and C, molecules non-marked), we identified seven other NICEdrug.ch molecules that clustered tightly with these statins (Figure 6.6 panel B and C, molecules marked with *). These new molecules share the same reactive site and physicochemical properties, and they have the highest similarity with known statins in atoms neighboring the reactive site. In a previous study by Endo *et al.*, these seven NICEdrug.ch molecules were introduced as Mevastatin analogues for inhibiting cholesterol biosynthesis [56]. Therefore, they were already suggested as possible candidates for treating high blood cholesterol and could be a good option for repurposing. Furthermore, we found eight known drugs not from the statin family among the 254 scanned molecules (appendix, Table 8.25). One of them, acetyl-L-carnitine (Figure 6.6 panel C, molecule marked with **), is mainly used for treating neuropathic pain [57], though Tanaka *et al.* have already confirmed that it also has a cholesterol-reducing effect [58].

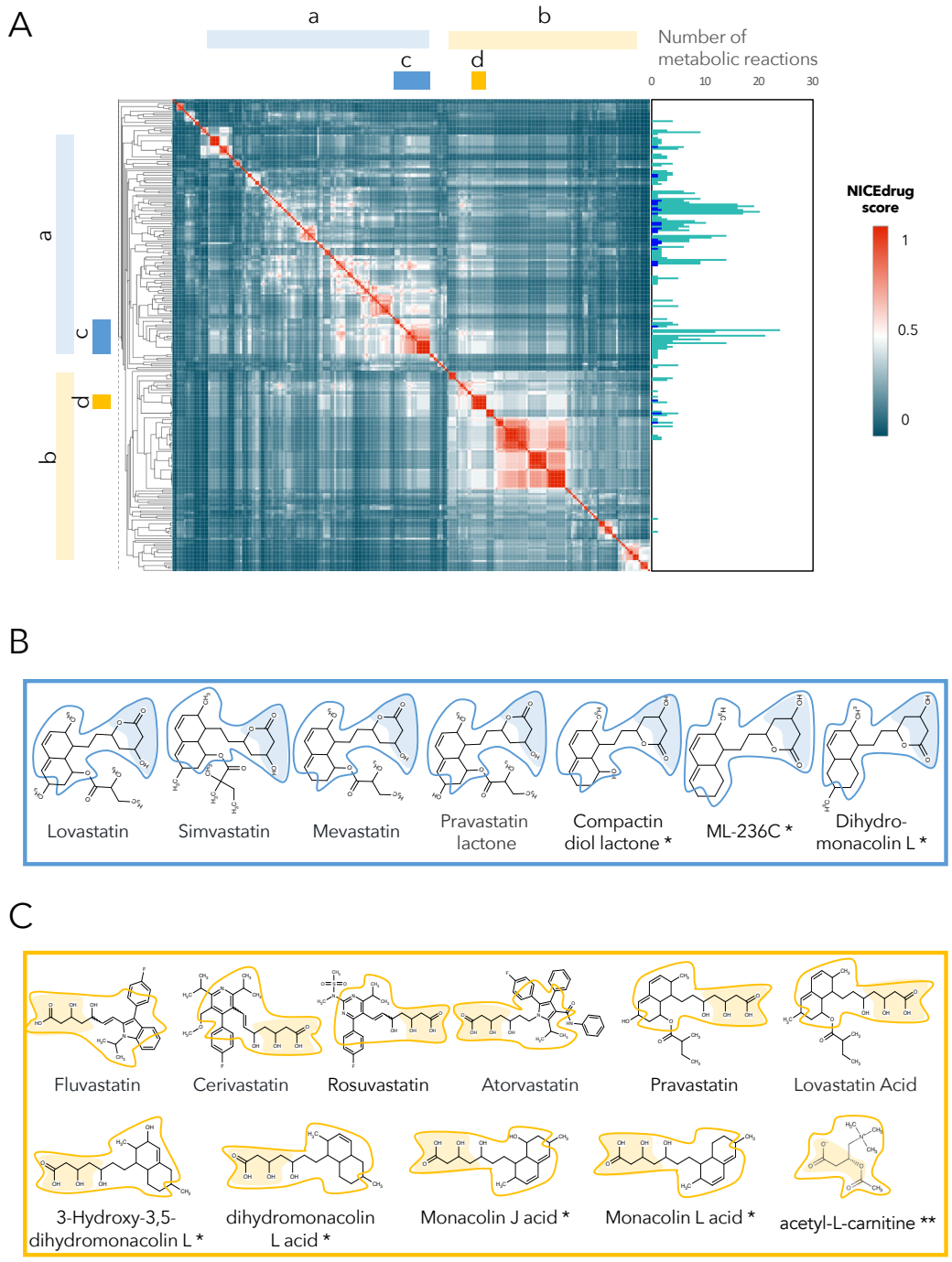


Figure 6.6: Clustering of molecules with statin reactive sites based on NICEdrug score suggests drugs for repurposing.

(A) Pairwise NICEdrug score between all molecules with statin reactive sites (heat map) and number of metabolic reactions in which they participate (right). We highlight clusters of statins of type 1 (cluster a) and type 2 (cluster b), and clusters of most similar molecules to type 1 statins (cluster c) and type 2 statins (cluster d). Within the metabolic reactions, we indicate the total number of reactions (dark color) and the number of reactions that involve the statin reactive site (light color). (B) Examples of statins and Mevastatin analogues of type 1 from cluster c (blue) and of type 2 from cluster d (gold). We left the known statins unmarked, which are appropriately clustered together based on the NICEdrug score, and we mark with * new molecules that cluster with statins and

that NICEdrug.ch suggests could be repurposed to act as statins. Reactive sites in type 1 statins and type 2 statins are colored in blue and orange, respectively. The reactive site neighborhood as considered in the NICEdrug score is also marked

Overall, NICEdrug.ch was able to characterize all known enzymatic reactions that metabolize statins, including proposed alternatives and new hypothetical reactions that could be involved in their metabolism within human cells Figure 6.6 panel A, Appendix Figure 8.5). The identification of seven drugs that clustered around the statins and were already designed as alternatives to statins verifies the ability of NICEdrug.ch and the NICEdrug score to search broad databases for similar compounds in structure and function. Furthermore, the discovery of the eight compounds unrelated to known statins offer multiple candidate repurposable drugs along with a map of their metabolized intermediates for the treatment of high cholesterol, though further preclinical experiments would be required to verify their clinical benefits.

6.3.5 NICEdrug.ch suggests over 500 drugs to target liver-stage malaria and simultaneously minimize side effects in human cells, with shikimate 3-phosphate as a top candidate

Efficiently targeting malaria remains a global health challenge. Malaria parasites (*Plasmodium*) are developing resistance to all known drugs, and antimalarials cause many side effects [59]. We applied NICEdrug.ch to identify drug candidates that target liver-stage developing malaria parasites and lessen or avoid side effects in human cells.

We previously reported 178 essential genes and enzymes for liver-stage development in the malaria parasite *Plasmodium berghei* [32] (appendix, Table 8.26). Out of 178 essential *Plasmodium* enzymes, 32 enzymes are not essential in human cells [34] (appendix, Table 8.26). We extracted all molecules catalyzed by these 32 enzymes uniquely essential in *Plasmodium*, which resulted in 68 metabolites and 157 unique metabolite-enzyme pairs (appendix, Table 8.26). We used NICEdrug.ch to examine the druggability of the 32 essential *Plasmodium* enzymes with the curated 48,544 drugs (Figure 6.1) and the possibility of repurposing them to target malaria.

We considered as candidates for targeting liver-stage malaria as the drugs or their metabolic neighbors that show a good NICEdrug score (NICEdrug score above 0.5) with any of the 157 *Plasmodium* metabolite-enzyme pairs. We identified 516 such drug candidates, targeting 16 essential *Plasmodium* enzymes (appendix, Table 8.27). Furthermore, 1,164 other drugs appear in the metabolic neighborhood of the 516 identified drugs (between one and three reaction steps away). Interestingly, out of the 516 identified drug candidates, digoxigenin, estradiol-17beta and estriol have been previously validated as antimalarials [60] and NICEdrug.ch suggests their antimalarial activity relies on the competitive inhibition of the KRC enzyme (Figure 6.7). This enzyme is part of both the steroid metabolism and the fatty acid elongation metabolism, which we

recently showed is essential for *Plasmodium* liver-stage development [32]. Among the 516 NICEdrug antimalarial candidates, there are also 89 molecules present in the metabolic neighborhood of antimalarial drugs approved by [60], which suggests these antimalarials might be prodrugs (appendix, Table 8.27).

Being an intracellular parasite, antimalarial treatments should be efficient at targeting *Plasmodium* as well as assure the integrity of the host cell (Figure 6.7 panel A). To tackle this challenge, we identified 1,497 metabolites participating in metabolic reactions catalyzed with essential human enzymes (appendix, Table 8.26, see 6.2.13) and excluded the antimalarial drug candidates that shared reactive site-centric similarity with the extracted human metabolite set (to satisfy NICEdrug score below 0.5). Out of all 516 drug candidates that might target liver-stage *Plasmodium*, a reduced set of 64 molecules minimize the inhibition of essential human enzymes (appendix, Table 8.27, see 6.2.14) and are hence optimal antimalarial candidates.

Among our set of 64 optimal antimalarial candidates, a set of 14 drugs targeting the *Plasmodium* shikimate metabolism, whose function is essential for liver-stage malaria development [32], arose as the top candidate because of its complete absence in human cells. The set of drugs targeting shikimate metabolism include 40 prodrugs (between one and three reaction steps away) that have been shown to have antimalarial activity [60] (appendix, Table 8.27). NICEdrug.ch identified molecules among the prodrugs with a high number of toxic alerts, like nitrofen. It also identified four molecules with scaffolds similar (two or three steps away) to the 1-(4-chlorobenzoyl)pyrazolidin-3-one of shikimate and derivatives. This result suggests that downstream compounds of the 40 prodrugs might target the *Plasmodium* shikimate pathway, but also might cause side effects in humans (appendix, Table 8.27).

To this end, NICEdrug.ch identified shikimate 3-phosphate as a top candidate antimalarial drug. We propose that shikimate 3-phosphate inhibits the essential *Plasmodium* shikimate biosynthesis pathway without side effects in the host cell (Figure 6.7, appendix, Table 8.27). Excitingly, shikimate 3-phosphate has been used to treat *E. coli* and *Streptococcus* infections without appreciable toxicity for patients [61]. Furthermore, recent studies have shown that inhibiting the shikimate pathway using 7-deoxy-sedoheptulose is an attractive antimicrobial and herbicidal strategy with no cytotoxic effects on mammalian cells [62]. Experimental studies should now validate the capability of shikimate 3-phosphate to efficiently and safely target liver malaria, and could further test other NICEdrug.ch antimalarial candidates (appendix, Table 8.27).

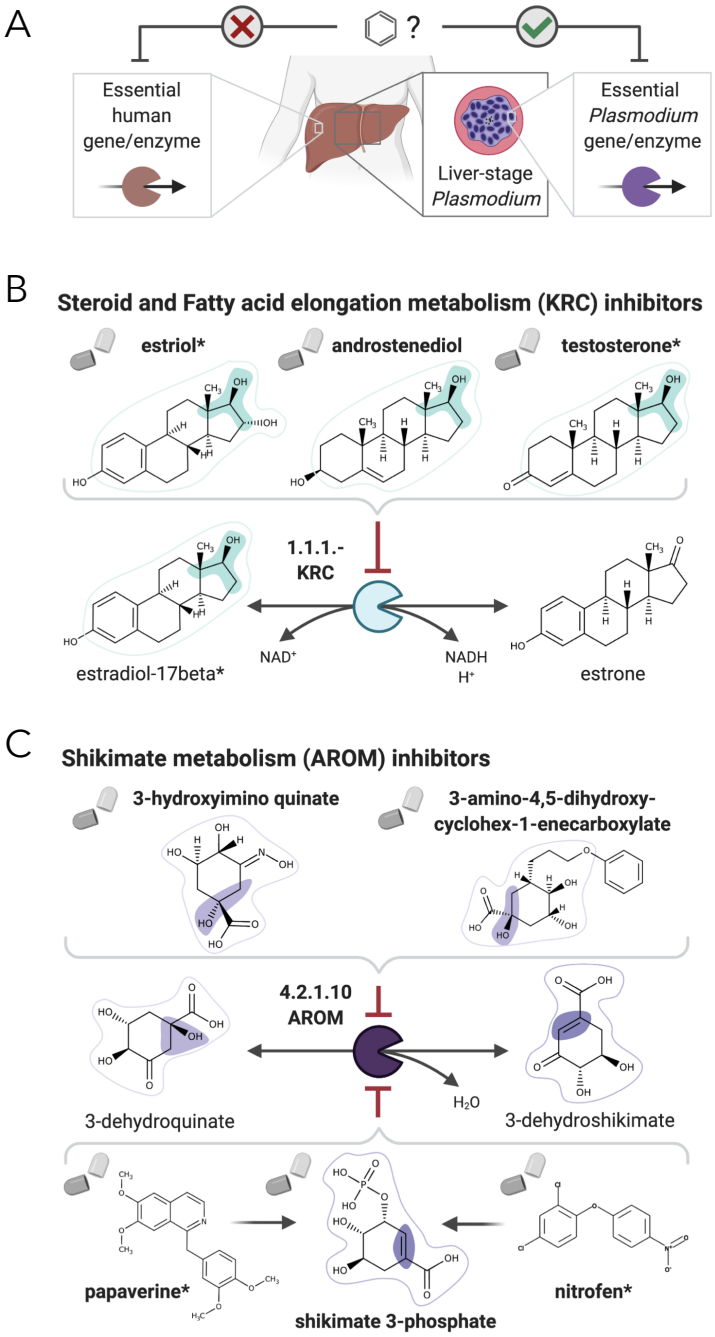


Figure 6.7: NICEdrug.ch suggests shikimate 3-phosphate as a top candidate to target liver-stage malaria and minimize side effects in host human cells. (A) Schema of ideal scenario to target malaria, wherein a drug efficiently inhibits an essential enzyme for malaria parasite survival and does not inhibit essential enzymes in the host human cell to prevent side effects. (B) Shikimate 3-phosphate inhibits enzymes in the *Plasmodium* shikimate metabolism, which is essential for liver-stage development of the parasite. Shikimate 3-phosphate does not inhibit any enzyme in the human host cell since it is not a native human metabolite, and it does not show similarity to any native human metabolite. (C) Mechanistic details of inhibition of aroC by shikimate 3-phosphate and other NICEdrug candidates

NICEdrug.ch identifies over 1,300 molecules to fight COVID-19, with N-acetylcysteine as a top candidate

SARS-CoV-2 is responsible for the currently on-going COVID-19 pandemic and the death of over half a million people (as of today, July 15 [63]) and there is currently no confirmed treatment for it. Attacking the host factors that allow replication and spread of the virus is an attractive strategy to treat viral infections like COVID-19. A recent study has identified 332 interactions between SARS-CoV-2 proteins and human proteins, which involve 332 hijacked human proteins or host factors [36]. Here, we first used NICEdrug.ch to identify inhibitors of enzymatic host factors of SARS-CoV-2. Targeting such human enzymes prevents interactions between human and viral proteins (PPI) (See section 6.2.16, Figure 6.8 panel A). Out of the 332 hijacked human proteins we identified 97 enzymes (See section 6.2.16, appendix, Table 8.28) and evaluated their druggability by inhibitors among the 250,000 small molecules in NICEdrug.ch and 80,000 molecules in food (See section 6.2.15, Figure 6.8 panel A). NICEdrug.ch suggests 22 hijacked human enzymes can be drug targets, and proposed 1301 potential competitive inhibitors from the NICEdrug.ch database. Out of 1301 potential inhibitors, 465 are known drugs, 712 are active metabolic products of 1,419 one-step-away prodrugs, and 402 are molecules in fooDB (appendix, Table 8.28). We found among the top anti SARS-CoV-2 drug candidates the known reverse transcriptase inhibitor didanosine (Figure 6.8 panel B, appendix, Table 8.28), which other *in silico* screenings have also suggested as a potential treatment for COVID-19 [64], [65]. Among others, NICEdrug.ch also identified: (1) actodigin, which belongs to the family of cardiotonic molecules proven to be effective against MERS-CoV but without mechanistic knowledge [66], (2) three molecules in ginger (6-paradol, 10-gingerol, and 6-shogaol) inhibiting catechol methyltransferase, and (3) brivudine, a DNA polymerase inhibitor that has been used to treat herpes zoster [67] and prevent MERS-CoV infection [68], and NICEdrug.ch suggests it for repurposing (Figure 6.9, appendix, Table 8.28).

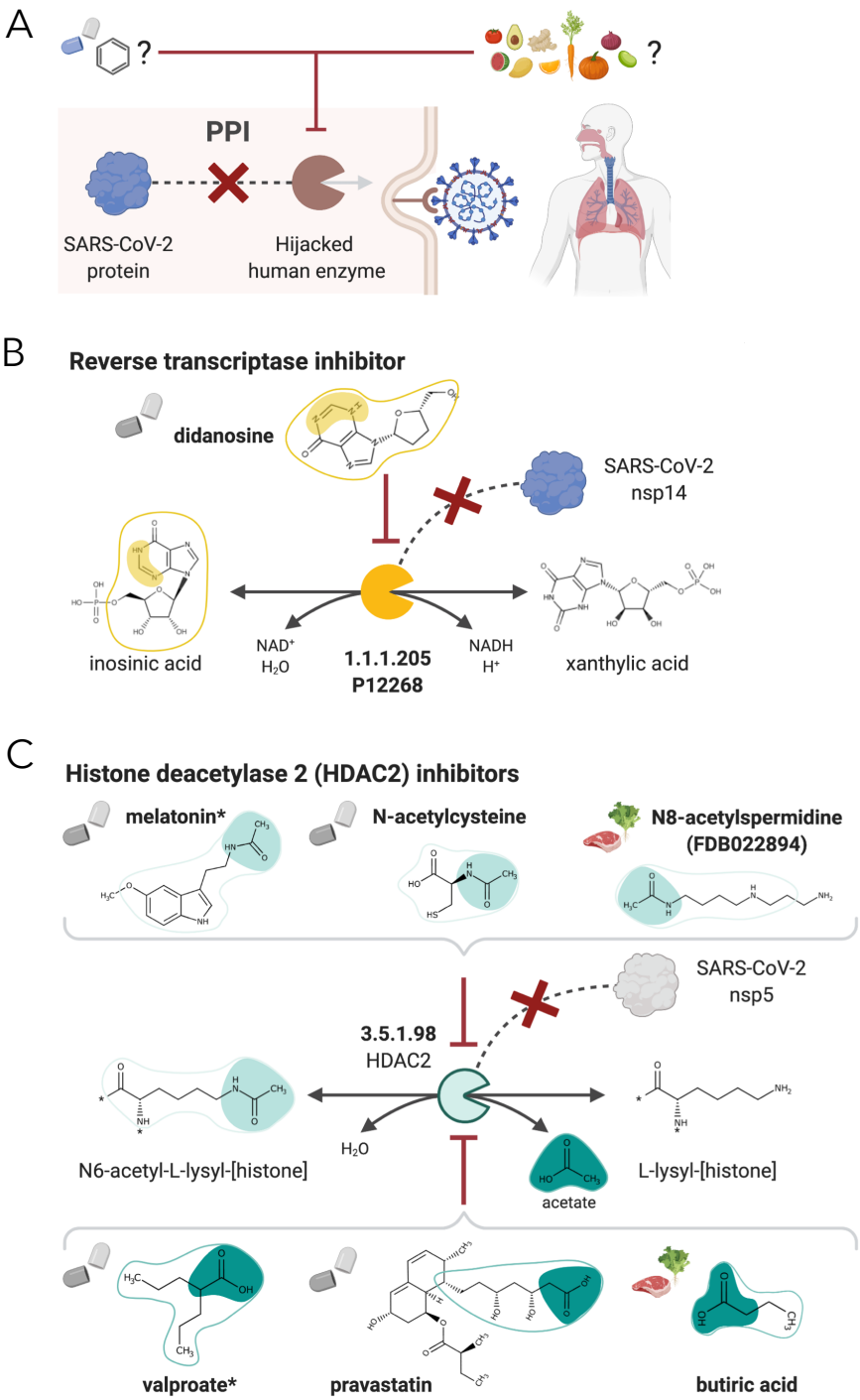


Figure 6.8: NICEdrug strategy to fight COVID-19, and NICEdrug candidate inhibitors of SARS-CoV-2 host factors: reverse transcriptase and HDAC2. (A) Schema of NICEdrug strategy to target COVID-19, wherein a drug (top-left) or molecules in food (top-right) efficiently inhibit a human enzyme hijacked by SARS-CoV-2. Inhibition of this host factor reduces or abolishes protein-protein interactions (PPI) with a viral protein and prevents SARS-CoV-2 proliferation. (B) Inhibition of the reverse transcriptase (E.C: 1.1.1.205 or P12268) and the PPI with SARS-CoV-nsp14 by didanosine based on NICEdrug.ch. (C) Inhibition of the HDAC2 (E.C: 3.5.1.98) and the PPI with SARS-

CoV-nsp5 by molecules containing acetyl moiety (like melatonin, N-acetylcysteine, and N8-acetylspermidine), and molecules containing carboxylate moiety (like valproate, stains, and butyrate) based on NICEdrug.ch

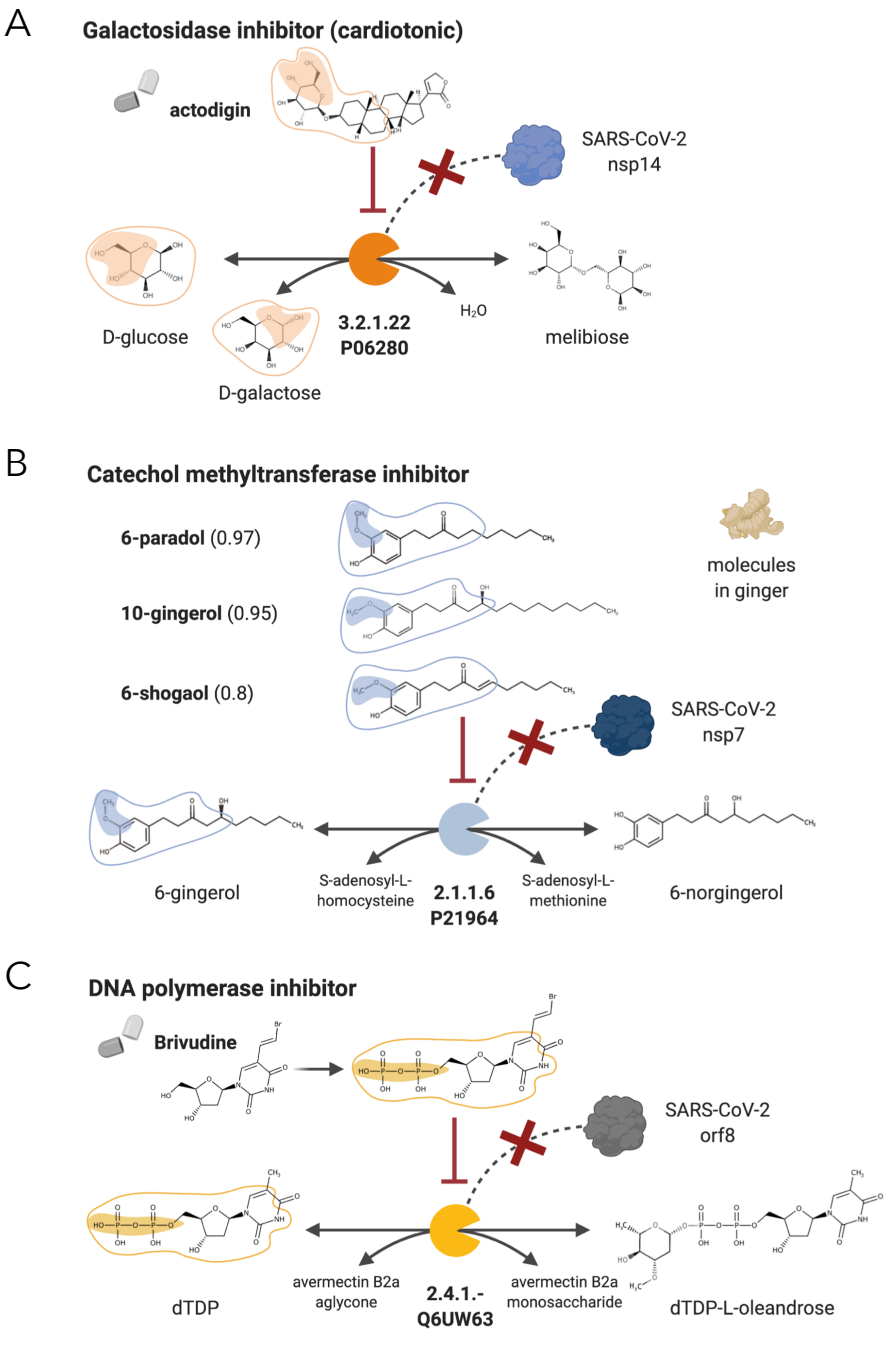


Figure 6.9: NICEdrug candidate inhibitors of SARS-CoV-2 host factors: galactosidase, catechol methyltransferase, and DNA polymerase, related to Figure 6.8. (A) Inhibition of the galactosidase (E.C: 3.2.1.22 or P06280) and the PPI with SARS-CoV-2 nsp14 by actodigin based on NICEdrug.ch. (B) Inhibition of the catechol methyltransferase (E.C: 2.1.1.6 or P21964) and the PPI with SARS-CoV-

2 nsp7 by 6-paradol, 10-gingerol, and 6-shogaol, which are molecules in ginger, based on NICEdrug.ch. (C) Inhibition of the DNA polymerase (E.C: 2.4.1.-) and the PPI with SARS-CoV-2 nsp8 by brivudine based on NICEdrug.ch.

Drugs like remdesivir, EIDD-2801, favipiravir, and inhibitors of angiotensin converting enzyme 2 (ACE2) have been used to treat COVID-19 [69], and act through a presumably effective inhibitory mechanism (Figure 6.10). For instance, the three drugs remdesivir, EIDD-2801, and favipiravir are believed to inhibit the DNA-directed RNA polymerase (E.C: 2.7.7.6). Here, we used the NICEdrug reactive site-centric fingerprint to seek for alternative small molecules in NICEdrug.ch and fooDB that could be repurposed to target ACE2 and DNA-directed RNA polymerase. NICEdrug.ch identified a total of 215 possible competitive inhibitors of ACE2. Among those is captopril, a known ACE2 inhibitor [70], and D-leucyl-N-(4-carbamimidoylbenzyl)-L-prolinamide, a NICEdrug.ch suggestion for drug repurposing to treat COVID-19. We also found 39 food-based molecules with indole-3-acetyl-proline (a molecule in soybean) as top ACE2 inhibitor candidate (Figure 6.10 , appendix, Table 8.29). To target the same enzyme as remdesivir, EIDD-2801, and favipiravir, NICEdrug.ch identified 1115 inhibitors of the DNA-directed RNA polymerase, like the drug vidarabine, which shows broad spectrum activity against DNA viruses in cell cultures and significant antiviral activity against infections like the herpes viruses, the vaccinia virus, and varicella zoster virus [71]. We further found 556 molecules in food that might inhibit DNA-directed RNA polymerase, like trans-zeatin riboside triphosphate (FDB031217) (appendix, Table 8.29).

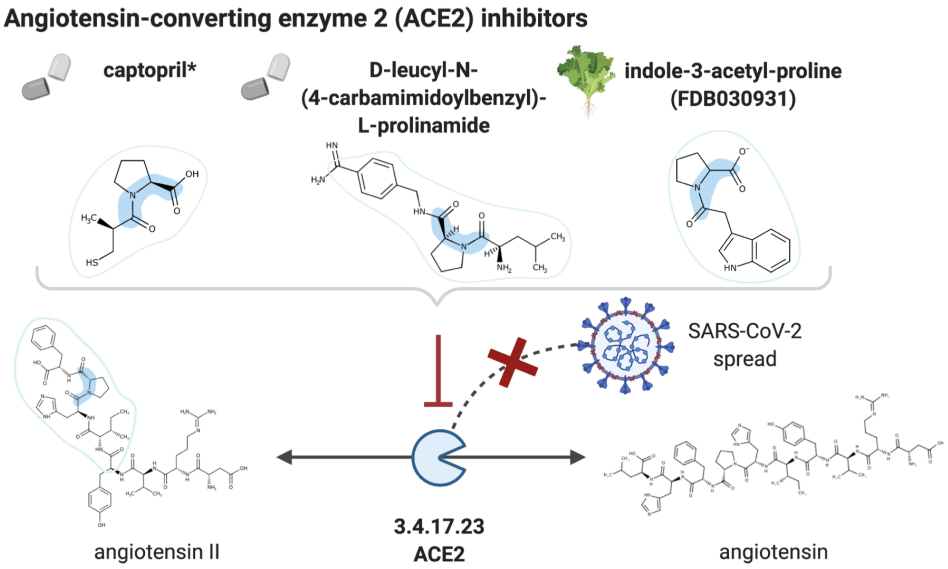


Figure 6.10: NICEdrug candidate inhibitors of ACE2, related to Figure 6.8. Inhibition of the ACE2 (E.C: 3.4.17.23), a putative host factor of SARS-CoV-2, by the known inhibitor captopril, and NICEdrug candidates D-leucyl-N-(4-carbamimidoylbezyl)-L-prolinamide and indole-3-acetyl-proline.

One of the host factors identified by Gordon and co-workers is the histone deacetylase 2 (HDAC2) [36], which acetylates proteins and is an important transcriptional and epigenetic regulator. The acetyl and carboxylate moieties are the reactive sites of the forward (N6-acetyl-L-lysyl-[histone]) and reverse (acetate) biotransformation of HDAC2, respectively (Figure 6.8). NICEdrug.ch recognized a total of 640 drugs for repurposing that can inhibit HDAC2, including 311 drugs sharing the acetyl moiety and showing a NICEdrug score above 0.5 with respect to N6-acetyl-L-lysyl-[histone], and 329 drugs sharing the carboxylate moiety and presenting a NICEdrug score above 0.5 with acetate (See 6.2.6). Among the drugs sharing the acetyl reactive site, we identified the known HDAC2 inhibitor melatonin [72], and to-our-knowledge new candidates like N-acetylhistamine and N-acetylcysteine. We also located 22 molecules in food with potential HDAC2 inhibitory activity, like N8-acetylspermidine (FDB022894) (Figure 6.8 panel C, appendix, Table 8.29). Drugs sharing the carboxylate reactive site (as identified with NICEdrug) include the known HDAC2 inhibitors valproate, butyrate, phenyl butyrate [73] and statins [52] (Figure 6.8 panel C, appendix, Table 8.29). Interestingly, statins have been shown to have protective activity against SARS-CoV-2 [74], [75]. In addition and excitingly, the NICEdrug.ch candidate N-acetylcysteine is a commonly used mucolytic drug that is sometimes considered as a dietary supplement and has putative antioxidant properties. Indeed, N-acetylcysteine is believed for long to be precursor of the cellular antioxidant glutathione [76], but has unknown pharmacological action. NICEdrug.ch suggests that N-acetylcysteine might present a dual antiviral activity: firstly, N-acetylcysteine is converted to cysteine by HDAC2 and by that means, it is competitively inhibiting the native function of HDAC2 and interactions with viral proteins (Figure 6.8 panel C, appendix, Table 8.29). Cysteine next fuels the glutathione biosynthesis pathway and produces glutathione in two steps.

Given the high coverage of validated molecules with activity against SARS-CoV-2 that NICEdrug.ch captured in this unbiased and reactive site-centric analysis, we suggest there might be other molecules in the set of 1,300 NICEdrug.ch candidates that could also fight COVID-19. Excitingly, there are many molecules that can be directly tested since these are drugs that have already passed all safety regulations or are molecules in food, like N-acetylcysteine for which we further reveal an action mechanism behind its potential anti SARS-CoV-2 activity. Other new candidates for which no safety data is available should be further validated experimentally and clinically. The mechanistic analyses provided by NICEdrug.ch could also guide new pharmacokinetic and pharmacodynamic models simulating SARS-CoV-2 infection and treatment.

6.4 Conclusion and outlook

To systematically illuminate the metabolism and all enzymatic targets (competitively inhibited) of known drugs and hypothetical prodrugs to aid in the development of new therapeutic compounds, we used a proven reaction-prediction tool BNICE.ch [12] and an analysis of neighboring atoms of reactive sites analogous to BridgIT [7] and performed the first large-scale computational analysis of drug biochemistry and toxicity in the context of human metabolism. The analysis involved over 250,000 small molecules, and curation and computation of bio- and physicochemical drug properties that we assembled in an open-source drug database NICEdrug.ch that can generate detailed drug metabolic reports and can be easily accessed and used by researchers, clinicians, and industry partners. Excitingly, NICEdrug.ch revealed 20 million potential reactive sites at the 250,000 small molecules of the database, and there exist over 3,000 enzymes in the human metabolism that can be inhibited with the 250,000 molecules. This is because NICEdrug.ch can identify *all* potential metabolic intermediates of a drug and scans these molecules for substructures that can interact with catalytic sites across all enzymes in a desired cell.

NICEdrug.ch adapts the metric previously developed for reactions in BridgIT [7] to precisely compare drug-drug and drug-metabolite pairs based on similarity of reactive site and the neighborhood around this reactive site, which we have recently shown outperforms previously defined molecular comparison metrics [7]. Since NICEdrug.ch shows high specificity in the identification of such reactive sites and neighborhood, it provides a better mechanistic understanding than currently available methods [77]. Despite these advances, it remains challenging to systematically identify non-competitive inhibition or targeting of non-enzymatic biological processes. We suggest coupling NICEdrug.ch drug metabolic reports with other *in silico* and experimental analyses accounting for signaling induction of small molecules and other non-enzymatic biological processes like transport of metabolites in a cell. The combined analysis of drug effects on different possible biological targets (not uniquely enzymes) will ultimately increase the coverage of molecules for which a mechanistic understanding of their mode of action is assigned.

A better understanding of the mechanisms of interactions and the specific nodes where the compounds act can help re-evaluate pharmacokinetic and pharmacodynamic models, dosage, and treatment. Such understanding can be used in the future to build models that correlate the pharmacodynamic information with specific compounds and chemical substructures in a manner similar to the one used for correlating compound structures with transcriptomic responses. We have shown for one of the most commonly used anticancer drugs, 5-FU, that NICEdrug.ch identifies and ranks alternative sources of toxicity and hence can guide the design of updated models and treatments to alleviate the drug's side-effects.

The mechanistic understanding will also further promote the development of drugs for repurposing. While current efforts in repurposing capitalize on the accepted status of known drugs, some of the issues with side

effects and unknown interactions limit their development as drugs for new diseases. Given that drug repurposing will require new dosage and administration protocols, the understanding of their interactions with the human metabolism will be very important in identifying, developing, and interpreting unanticipated side effects and physiological responses. We evaluated the possibility of drug repurposing with NICEdrug.ch as a substitute for statins, which are broadly used to reduce cholesterol but have many side effects. NICEdrug.ch and its reactive site-centric comparison accurately cluster both family types of statins, even though they are similar in overall molecular structure and show different reactivity. In addition, NICEdrug.ch suggests a set of new molecules with hypothetically less side effects [56], [58] that share reactive sites with statins.

A better mechanistic understanding of drug targets can guide the design of treatments against infectious diseases, for which we need effective drugs that target pathogens without side effects in the host cell. This is arguably the most challenging type of problem in drug design, and indeed machine learning has continuously failed to guide such designs given the difficulty in quantifying side effects—not to mention in acquiring large, consistent, and high-quality data sets from human patients. To demonstrate the power of NICEdrug.ch for tackling this problem, we sought to identify drugs that target liver-stage malaria parasites and minimize the impact on the human host cell. We identified over 500 drugs that inhibit essential *Plasmodium* enzymes in the liver stages and minimize the impact on the human host cell. Our top drug candidate is shikimate 3-phosphate targeting the parasite's shikimate metabolism, which we recently identified as essential in a high-throughput gene knockout screening in *Plasmodium* [32]. Excitingly, our suggested antimalarial candidate shikimate 3-phosphate has already been used for *Escherichia* and *Streptococcus* infections without appreciable side effects [61].

Finally, minimizing side effects becomes especially challenging in the treatment of viral infections, since viruses fully rely on the host cell to replicate. As a last demonstration of the potential of NICEdrug.ch, we sought to target COVID-19 by identifying inhibitors of 22 known enzymatic host factors of SARS-CoV-2 [36]. NICEdrug.ch identified over 1,300 molecules that might target the 22 host factors and prevent SARS-CoV-2 replication. As a validation, NICEdrug.ch correctly identified known inhibitors of those enzymes, and further suggested safe drugs for repurposing and other food molecules with activity against SARS-CoV-2. Among the NICEdrug.ch suggestions for COVID-19, based on the knowledge on its mechanism and safety, we highlight N-acetylcysteine as an inhibitor of HDAC2 and SARS-CoV-2.

Overall, we believe that a systems level or metabolic network analysis coupled with an investigation of reactive sites will likely accelerate the discovery of new drugs and provide additional understanding regarding metabolic fate, action mechanisms, and side effects and can complement on-going experimental effects to understand drug metabolism [8]. We suggest the generation of drug metabolic reports to understand the reactivity of new small molecules, the possibility of drug repurposing, and the druggability of enzymes. Our

results using NICEdrug.ch suggest that this database can be a novel avenue towards the systematic pre-screening and identification of drugs and antimicrobials. In addition to human metabolic information, NICEdrug.ch currently includes information for the metabolism of *P. berghei* and *E. coli*. Because we are making it publicly available (https://lcsb-databases.epfl.ch/pathways/Nicedrug/), our hope is that scientists and medical practitioners alike can make use of this unique database to better inform their research and clinical decisions—saving time, money, and ultimately lives.

6.5 References

[1] C. H. Wong, K. W. Siah, and A. W. Lo, "Estimation of clinical trial success rates and related parameters," *Biostatistics*, vol. 20, no. 2, pp. 273–286, Apr. 2019, doi: 10.1093/biostatistics/kxx069.

[2] S. Shilo, H. Rossman, and E. Segal, "Axes of a revolution: challenges and promises of big data in healthcare," *Nat. Med.*, vol. 26, no. 1, pp. 29–38, Jan. 2020, doi: 10.1038/s41591-019-0727-5.

[3] J. M. Stokes *et al.*, "A Deep Learning Approach to Antibiotic Discovery," *Cell*, vol. 180, no. 4, pp. 688-702.e13, Feb. 2020, doi: 10.1016/j.cell.2020.01.021.

[4] J. Vamathevan *et al.*, "Applications of machine learning in drug discovery and development," *Nat. Rev. Drug Discov.*, vol. 18, no. 6, pp. 463–477, Jun. 2019, doi: 10.1038/s41573-019-0024-5.

[5] A. Jarvis and G. Ouvry, "Essential ingredients for rational drug design," *Bioorg. Med. Chem. Lett.*, vol. 29, no. 20, p. 126674, Oct. 2019, doi: 10.1016/j.bmcl.2019.126674.

[6] C. L. Verlinde and W. G. Hol, "Structure-based drug design: progress, results and challenges," *Structure*, vol. 2, no. 7, pp. 577–587, Jul. 1994, doi: 10.1016/S0969-2126(00)00060-5.

[7] N. Hadadi, H. MohammadiPeyhani, L. Miskovic, M. Seijo, and V. Hatzimanikatis, "Enzyme annotation for orphan and novel reactions using knowledge of substrate reactive sites," *Proc. Natl. Acad. Sci.*, p. 201818877, Mar. 2019, doi: 10.1073/pnas.1818877116.

[8] B. Javdan *et al.*, "Personalized Mapping of Drug Metabolism by the Human Gut Microbiome.," *Cell*, Jun. 2020, doi: 10.1016/j.cell.2020.05.001.

[9] N. H. Lim *et al.*, "Reactive-site mutants of N-TIMP-3 that selectively inhibit ADAMTS-4 and ADAMTS-5: biological and structural implications," *Biochem. J.*, vol. 431, no. 1, pp. 113–122, Oct. 2010, doi: 10.1042/BJ20100725.

[10] M. A. Ghattas, N. Raslan, A. Sadeq, M. Al Sorkhy, and N. Atatreh, "Druggability analysis and classification of protein tyrosine phosphatase active sites," *Drug Des. Devel. Ther.*, vol. 10, pp. 3197–3209, Sep. 2016, doi: 10.2147/DDDT.S111443.

[11] K. C. Soh and V. Hatzimanikatis, “DREAMS of metabolism,” *Trends Biotechnol.*, vol. 28, no. 10, pp. 501–508, Oct. 2010, doi: 10.1016/j.tibtech.2010.07.002.

[12] V. Hatzimanikatis, C. Li, J. A. Ionita, C. S. Henry, M. D. Jankowski, and L. J. Broadbelt, “Exploring the diversity of complex metabolic networks,” *Bioinformatics*, vol. 21, no. 8, pp. 1603–1609, Apr. 2005, doi: 10.1093/bioinformatics/bti213.

[13] N. Hadadi, J. Hafner, A. Shajkofci, A. Zisaki, and V. Hatzimanikatis, “ATLAS of Biochemistry: A Repository of All Possible Biochemical Reactions for Synthetic Biology and Metabolic Engineering Studies,” *ACS Synth. Biol.*, vol. 5, no. 10, pp. 1155–1166, Oct. 2016, doi: 10.1021/acssynbio.6b00054.

[14] J. Hafner, H. MohammadiPeyhani, A. Sveshnikova, A. Scheidegger, and V. Hatzimanikatis, “Updated ATLAS of Biochemistry with New Metabolites and Improved Enzyme Prediction Power,” *ACS Synth. Biol.*, vol. 9, no. 6, pp. 1479–1482, Jun. 2020, doi: 10.1021/acssynbio.0c00052.

[15] A. Scalbert *et al.*, “Databases on food phytochemicals and their health-promoting effects.,” *J. Agric. Food Chem.*, vol. 59, no. 9, pp. 4331–4348, May 2011, doi: 10.1021/jf200591d.

[16] S. D. Finley, L. J. Broadbelt, and V. Hatzimanikatis, “Computational Framework for Predictive Biodegradation,” *Biotechnol. Bioeng.*, vol. 104, no. 6, pp. 1086–1097, Dec. 2009, doi: 10.1002/bit.22489.

[17] N. Hadadi and V. Hatzimanikatis, “Design of computational retrobiosynthesis tools for the design of de novo synthetic pathways,” *Curr. Opin. Chem. Biol.*, vol. 28, pp. 99–104, Oct. 2015, doi: 10.1016/j.cbpa.2015.06.025.

[18] C. S. Henry, L. J. Broadbelt, and V. Hatzimanikatis, “Discovery and analysis of novel metabolic pathways for the biosynthesis of industrial chemicals: 3-hydroxypropanoate,” *Biotechnol. Bioeng.*, p. n/a-n/a, 2010, doi: 10.1002/bit.22673.

[19] M. Tokic *et al.*, “Discovery and Evaluation of Biosynthetic Pathways for the Production of Five Methyl Ethyl Ketone Precursors,” *ACS Synth. Biol.*, vol. 7, no. 8, pp. 1858–1873, Aug. 2018, doi: 10.1021/acssynbio.8b00049.

[20] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, “Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings1PII of original article: S0169-409X(96)00423-1. The article was originally published in Advanced Drug Delivery Reviews 23 (1997) 3–25.1,” *Adv. Drug Deliv. Rev.*, vol. 46, no. 1, pp. 3–26, Mar. 2001, doi: 10.1016/S0169-409X(00)00129-0.

[21] D. Weininger, “SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules,” *J. Chem. Inf. Comput. Sci.*, vol. 28, no. 1, pp. 31–36, Feb. 1988, doi: 10.1021/ci00057a005.

[22] N. M. O’Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, “Open Babel: An open chemical toolbox,” *J. Cheminformatics*, vol. 3, no. 1, p. 33, Oct. 2011, doi: 10.1186/1758-2946-3-33.

[23] M. D. Jankowski, C. S. Henry, L. J. Broadbelt, and V. Hatzimanikatis, “Group Contribution Method for Thermodynamic Analysis of Complex Metabolic Networks,” *Biophys. J.*, vol. 95, no. 3, pp. 1487–1499, Aug. 2008, doi: 10.1529/biophysj.107.124784.

[24] E. Brunk *et al.*, “Recon3D: A Resource Enabling A Three-Dimensional View of Gene Variation in Human Metabolism,” *Nat. Biotechnol.*, vol. 36, no. 3, pp. 272–281, Mar. 2018, doi: 10.1038/nbt.4072.

[25] N. Pornputtapong, I. Nookaew, and J. Nielsen, “Human metabolic atlas: an online resource for human metabolism,” *Database J. Biol. Databases Curation*, vol. 2015, Jul. 2015, doi: 10.1093/database/bav068.

[26] D. Croft *et al.*, “Reactome: a database of reactions, pathways and biological processes,” *Nucleic Acids Res.*, vol. 39, no. Database issue, pp. D691–D697, Jan. 2011, doi: 10.1093/nar/gkq1018.

[27] D. S. Wishart *et al.*, “HMDB 4.0: the human metabolome database for 2018,” *Nucleic Acids Res.*, vol. 46, no. Database issue, pp. D608–D617, Jan. 2018, doi: 10.1093/nar/gkx1089.

[28] V. I. Levenshtein, “Binary Codes Capable of Correcting Deletions, Insertions and Reversals,” *Sov. Phys. Dokl.*, vol. 10, p. 707, Feb. 1966.

[29] G. Papadatos and N. Brown, “In silico applications of bioisosterism in contemporary medicinal chemistry practice,” *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, vol. 3, no. 4, pp. 339–354, 2013, doi: 10.1002/wcms.1148.

[30] E. J. Ariens, *Molecular Pharmacology V3: The Model of Action of Biology Active Compounds*. Elsevier, 2012.

[31] I. Sushko, E. Salmina, V. A. Potemkin, G. Poda, and I. V. Tetko, “ToxAlerts: a Web server of structural alerts for toxic chemicals and compounds with potential adverse reactions,” *J. Chem. Inf. Model.*, vol. 52, no. 8, pp. 2310–2316, Aug. 2012, doi: 10.1021/ci300245q.

[32] R. R. Stanway *et al.*, “Genome-Scale Identification of Essential Metabolic Processes for Targeting the Plasmodium Liver Stage,” *Cell*, vol. 179, no. 5, pp. 1112-1128.e26, Nov. 2019, doi: 10.1016/j.cell.2019.10.030.

[33] E. Bushell *et al.*, "Functional Profiling of a Plasmodium Genome Reveals an Abundance of Essential Genes," *Cell*, vol. 170, no. 2, pp. 260-272.e8, Jul. 2017, doi: 10.1016/j.cell.2017.06.030.

[34] T. Wang *et al.*, "Identification and characterization of essential genes in the human genome," *Science*, vol. 350, no. 6264, p. 1096, Nov. 2015, doi: 10.1126/science.aac7041.

[35] J. L. Robinson *et al.*, "An atlas of human metabolism," *Sci. Signal.*, vol. 13, no. 624, p. eaaz1482, Mar. 2020, doi: 10.1126/scisignal.aaz1482.

[36] D. E. Gordon *et al.*, "A SARS-CoV-2 protein interaction map reveals targets for drug repurposing.," *Nature*, Apr. 2020, doi: 10.1038/s41586-020-2286-9.

[37] D. B. Longley, D. P. Harkin, and P. G. Johnston, "5-Fluorouracil: mechanisms of action and clinical strategies," *Nat. Rev. Cancer*, vol. 3, no. 5, pp. 330–338, May 2003, doi: 10.1038/nrc1074.

[38] B. Testa, "Principles of Drug Metabolism," in *Burger's Medicinal Chemistry and Drug Discovery*, American Cancer Society, 2010, pp. 403–454.

[39] W. W. Ma *et al.*, "Emergency use of uridine triacetate for the prevention and treatment of life-threatening 5-fluorouracil and capecitabine toxicity," *Cancer*, vol. 123, no. 2, pp. 345–356, Jan. 2017, doi: 10.1002/cncr.30321.

[40] A. C. Sartorelli and D. G. Johns, *Antineoplastic and Immunosuppressive Agents: Part I*. Springer Science & Business Media, 2013.

[41] R. Matsuda *et al.*, "STUDIES OF METABOLITE-PROTEIN INTERACTIONS: A REVIEW," *J. Chromatogr. B Analyt. Technol. Biomed. Life. Sci.*, vol. 966, p. 48, Sep. 2014, doi: 10.1016/j.jchromb.2013.11.043.

[42] M. Kobayakawa and Y. Kojima, "Tegafur/gimeracil/oteracil (S-1) approved for the treatment of advanced gastric cancer in adults when given in combination with cisplatin: a review comparing it with other fluoropyrimidine-based therapies," *OncoTargets Ther.*, vol. 4, pp. 193–201, Nov. 2011, doi: 10.2147/OTT.S19059.

[43] J. H. Bielas, M. W. Schmitt, A. Icreverzi, N. G. Ericson, and L. A. Loeb, "Molecularly Evolved Thymidylate Synthase Inhibits 5-Fluorodeoxyuridine Toxicity in Human Hematopoietic Cells," *Hum. Gene Ther.*, vol. 20, no. 12, pp. 1703–1707, Dec. 2009, doi: 10.1089/hum.2009.053.

[44] F. P. Guengerich, "Mechanisms of Drug Toxicity and Relevance to Pharmaceutical Development," *Drug Metab. Pharmacokinet.*, vol. 26, no. 1, pp. 3–14, 2011.

[45] U. Schmidt *et al.*, “SuperToxic: a comprehensive database of toxic compounds,” *Nucleic Acids Res.*, vol. 37, no. Database, pp. D295–D299, Jan. 2009, doi: 10.1093/nar/gkn850.

[46] S. Thakkar, M. Chen, H. Fang, Z. Liu, R. Roberts, and W. Tong, “The Liver Toxicity Knowledge Base (LKTB) and drug-induced liver injury (DILI) classification for assessment of human liver injury,” *Expert Rev. Gastroenterol. Hepatol.*, vol. 12, no. 1, pp. 31–38, Jan. 2018, doi: 10.1080/17474124.2018.1383154.

[47] H. Yang, J. Li, Z. Wu, W. Li, G. Liu, and Y. Tang, “Evaluation of Different Methods for Identification of Structural Alerts Using Chemical Ames Mutagenicity Data Set as a Benchmark,” *Chem Res Toxicol*, p. 10, 2017.

[48] J. Krauß and F. Bracher, “Pharmacokinetic Enhancers (Boosters)—Escort for Drugs against Degrading Enzymes and Beyond,” *Sci. Pharm.*, vol. 86, no. 4, p. 43, Dec. 2018, doi: 10.3390/scipharm86040043.

[49] M. Mahmoudian and P. Rahimi-Moghaddam, “The anti-cancer activity of noscapine: a review.,” *Recent Patents Anticancer Drug Discov.*, vol. 4, no. 1, pp. 92–97, Jan. 2009, doi: 10.2174/157489209787002524.

[50] Rajesh, A. and C. International, *Medicinal plant biotechnology*. Cambridge, MA: CABI, 2011.

[51] US Preventive Services Task Force, “Statin Use for the Primary Prevention of Cardiovascular Disease in Adults: US Preventive Services Task Force Recommendation Statement,” *JAMA*, vol. 316, no. 19, pp. 1997–2007, Nov. 2016, doi: 10.1001/jama.2016.15450.

[52] W. Kong *et al.*, “Berberine is a novel cholesterol-lowering drug working through a unique mechanism distinct from statins,” *Nat. Med.*, vol. 10, no. 12, pp. 1344–1351, Dec. 2004, doi: 10.1038/nm1135.

[53] S.-Y. Jiang *et al.*, “Discovery of a potent HMG-CoA reductase degrader that eliminates statin-induced reductase accumulation and lowers cholesterol,” *Nat. Commun.*, vol. 9, no. 1, p. 5138, Dec. 2018, doi: 10.1038/s41467-018-07590-3.

[54] F. Mulhaupt *et al.*, “Statins (HMG-CoA reductase inhibitors) reduce CD40 expression in human vascular cells,” *Cardiovasc. Res.*, vol. 59, no. 3, pp. 755–766, Sep. 2003, doi: 10.1016/S0008-6363(03)00515-7.

[55] E. S. Istvan, “Structural Mechanism for Statin Inhibition of HMG-CoA Reductase,” *Science*, vol. 292, no. 5519, pp. 1160–1164, May 2001, doi: 10.1126/science.1059344.

[56] A. Endo and K. Hasumi, “HMG-CoA reductase inhibitors,” *Nat. Prod. Rep.*, vol. 10, no. 6, p. 541, 1993, doi: 10.1039/np9931000541.

[57] S. Li, Q. Li, Y. Li, L. Li, H. Tian, and X. Sun, “Acetyl-L-Carnitine in the Treatment of Peripheral Neuropathic Pain: A Systematic Review and Meta-Analysis of Randomized Controlled Trials,” *PLoS ONE*, vol. 10, no. 3, Mar. 2015, doi: 10.1371/journal.pone.0119479.

[58] Y. Tanaka *et al.*, “Acetyl-L-carnitine supplementation restores decreased tissue carnitine levels and impaired lipid metabolism in aged rats,” *J. Lipid Res.*, vol. 45, no. 4, pp. 729–735, Apr. 2004, doi: 10.1194/jlr.M300425-JLR200.

[59] World Health Organization, “World Malaria Report 2018,” 9789241565653, 2018. [Online]. Available: www.who.int/malaria.

[60] Y. Antonova-Koch *et al.*, “Open-source discovery of chemical leads for next-generation chemoprotective antimalarials,” *Science*, vol. 362, no. 6419, p. eaat9446, Dec. 2018, doi: 10.1126/science.aat9446.

[61] D. C. Díaz-Quiroz *et al.*, “Synthesis, biological activity and molecular modelling studies of shikimic acid derivatives as inhibitors of the shikimate dehydrogenase enzyme of Escherichia coli,” *J. Enzyme Inhib. Med. Chem.*, vol. 33, no. 1, pp. 397–404, Jan. 2018, doi: 10.1080/14756366.2017.1422125.

[62] K. Brilisauer *et al.*, “Cyanobacterial antimetabolite 7-deoxy-sedoheptulose blocks the shikimate pathway to inhibit the growth of prototrophic organisms,” *Nat. Commun.*, vol. 10, no. 1, Art. no. 1, Feb. 2019, doi: 10.1038/s41467-019-08476-8.

[63] E. Dong, H. Du, and L. Gardner, “An interactive web-based dashboard to track COVID-19 in real time.,” *Lancet Infect. Dis.*, vol. 20, no. 5, pp. 533–534, May 2020, doi: 10.1016/S1473-3099(20)30120-1.

[64] F. M. Alakwaa, “Repurposing Didanosine as a Potential Treatment for COVID-19 Using Single-Cell RNA Sequencing Data.,” *mSystems*, vol. 5, no. 2, Apr. 2020, doi: 10.1128/mSystems.00297-20.

[65] C. Cava, G. Bertoli, and I. Castiglioni, “In Silico Discovery of Candidate Drugs against Covid-19,” *Viruses*, vol. 12, no. 4, p. 404, Apr. 2020, doi: 10.3390/v12040404.

[66] M. Ko *et al.*, “Screening of FDA-approved drugs using a MERS-CoV clinical isolate from South Korea identifies potential therapeutic options for COVID-19,” *bioRxiv*, p. 2020.02.25.965582, Jan. 2020, doi: 10.1101/2020.02.25.965582.

[67] S. Wassilew, “Brivudin compared with famciclovir in the treatment of herpes zoster: effects in acute disease and chronic pain in immunocompetent patients. A randomized, double-blind, multinational study.,” *J. Eur. Acad. Dermatol. Venereol. JEADV*, vol. 19, no. 1, pp. 47–55, Jan. 2005, doi: 10.1111/j.1468-3083.2004.01119.x.

[68] Y.-J. Park *et al.*, "Structures of MERS-CoV spike glycoprotein in complex with sialoside attachment receptors," *Nat. Struct. Mol. Biol.*, vol. 26, no. 12, pp. 1151–1157, Dec. 2019, doi: 10.1038/s41594-019-0334-7.

[69] S. Jeon *et al.*, "Identification of antiviral drug candidates against SARS-CoV-2 from FDA-approved drugs," *Antimicrob. Agents Chemother.*, p. AAC.00819-20, May 2020, doi: 10.1128/AAC.00819-20.

[70] H. M. Kim, D. R. Shin, O. J. Yoo, H. Lee, and J.-O. Lee, "Crystal structure of Drosophila angiotensin I-converting enzyme bound to captopril and lisinopril 1," *FEBS Lett.*, vol. 538, no. 1–3, pp. 65–70, Mar. 2003, doi: 10.1016/S0014-5793(03)00128-5.

[71] M. Suzuki, T. Okuda, and K. Shiraki, "Synergistic antiviral activity of acyclovir and vidarabine against herpes simplex virus types 1 and 2 and varicella-zoster virus.," *Antiviral Res.*, vol. 72, no. 2, pp. 157–161, Nov. 2006, doi: 10.1016/j.antiviral.2006.05.001.

[72] Y. Wu *et al.*, "The effect of melatonin on cardio fibrosis in juvenile rats with pressure overload and deregulation of HDACs," *Korean J. Physiol. Pharmacol. Off. J. Korean Physiol. Soc. Korean Soc. Pharmacol.*, vol. 22, no. 6, pp. 607–616, Nov. 2018, doi: 10.4196/kjpp.2018.22.6.607.

[73] M. M. Abdel-Atty, N. A. Farag, S. E. Kassab, R. A. T. Serya, and K. A. M. Abouzid, "Design, synthesis, 3D pharmacophore, QSAR, and docking studies of carboxylic acid derivatives as Histone Deacetylase inhibitors and cytotoxic agents," *Bioorganic Chem.*, vol. 57, pp. 65–82, Dec. 2014, doi: 10.1016/j.bioorg.2014.08.006.

[74] C. Lodigiani *et al.*, "Venous and arterial thromboembolic complications in COVID-19 patients admitted to an academic hospital in Milan, Italy," *Thromb. Res.*, vol. 191, pp. 9–14, Jul. 2020, doi: 10.1016/j.thromres.2020.04.024.

[75] X.-J. Zhang *et al.*, "In-hospital Use of Statins is Associated with a Reduced Risk of Mortality among Individuals with COVID-19," *Cell Metab.*, 2020, doi: 10.1016/j.cmet.2020.06.015.

[76] J. Mårtensson, J. Gustafsson, and A. Larsson, "A therapeutic trial with N-acetylcysteine in subjects with hereditary glutathione synthetase deficiency (5-oxoprolinuria).," *J. Inherit. Metab. Dis.*, vol. 12, no. 2, pp. 120–130, 1989, doi: 10.1007/BF01800713.

[77] J. G. Robertson, "Mechanistic Basis of Enzyme-Targeted Drugs," *Biochemistry*, vol. 44, no. 15, pp. 5561–5571, Apr. 2005, doi: 10.1021/bi050247e.

Chapter 7 **Conclusion and outlook**

"Biology adapted itself to the computer, not the computer to biology"

Hallam Stevens

The purpose of this final chapter is to review challenges entailed in this thesis and summarize the main findings. In addition, we discuss the future perspectives and the outline regarding computational modeling of metabolism and their potential applications in systems biology, metabolic engineering and drug discovery.

In this thesis, we employed computer programming in several disciplines such as chem-informatics, computational biology, bio-informatics, database development, etc. to explore the catalytic mechanism of enzymes. Understanding the structure and mechanism of enzymes with atom resolution has enabled us to mine, model and predict biochemistry. Here, we will recapitulate our major learnings in different chapters of this thesis.

To begin with, we need to know how to access, curate and implement data in an efficient way (chapter 2). The new advances in omics technology has led to accumulation of ever-growing amount of biological data and simultaneously the demand for innovative computational approaches for their analysis is increasing. High-performance, integrated databases that are capable of efficiently storing and searching for big amounts of biochemical data are demanding. Moreover, we need advanced algorithms for searching these databases. Towards this end, several databases were developed such as KEGG [1], MetaCyc [2], BRENDA [3], HMDB [4] that accommodate large portion of the reported biochemical data. Despite the recent efforts in developing biochemical databases, several challenges remain to be addressed: (i) the data entries in these databases are partially linked to each other, (ii) data exchange or data integration among several databases requires unification of datasets which is challenging, and (iii) the databases are typically focused on a specific organisms or pathway and at a specific scale. However, large-scale study of biological systems requires high quality homologous data covering vast range of biochemical networks. To handle this overload of big and heterogeneous data, a high level of data organization and data integration is necessary. An effective approach to integrate and classify large sums of data is to design and develop ontology-based databases. Ontology as a strategy for data classification, was developed in computer science to facilitate data reuse and data sharing [5]. It has been extensively used to model heterogeneous big data and the reason for this

popularity is due to its ability to keep semantics away from the type of data. In an ontology, each entity is defined a “concept” or an “instance” of a “concept” linked together by different relations. This structure in ontology allows to abstract data and capture the relationships between entries. Moreover, ontological design of a database allows for a flexible data organization. However, its application in biological and chemical data management and storage has not yet been fully explored. To overcome the problems of working with heterogeneous repositories, we developed an ontological database, called LCSB DB, and we integrated more than 14 external biochemical databases into our unified resource. LCSB DB enclosed biochemical data in several levels of compounds, enzymes, metabolic reactions, metabolic pathways and metabolic networks. To avoid replication inside database, we applied the established chem-informatics tools to convert compounds to several standardized formats such as canonical SMILES. Currently, the biochemical data integrated in LCSB DB accounts for 1M bio-compound and over 60k metabolic reactions (this section of LCSB DB is called bioDB). Further, by including chemicals and the results of computational tools developed in our group, e.g., BNICE.ch, number of compounds and reactions increase to more than 70M and over 5M respectively. LCSB DB serves as the standard platform to store and share data between users and computational tools. In future, integrating new data sources in LCSB DB is simply connecting the existing concepts and instances (or defining new ones) without the need to modify the architect of the database. Such an approach not only makes data integration easier and quicker, but also it standardizes the data definition and makes the semantic relationships among different entities explicit. In addition to regularly updating biochemical data in LCSB DB to keep up with the latest metabolic discoveries, we suggest expanding the currently defined properties. An example would be to annotate metabolic reactions with their kinetic characteristics. To do this, one should add kinetic rate laws as new concepts. Kinetic laws along with experimental conditions and parameter values will be linked as new properties to biochemical reactions. Further, we suggest expanding the scope of LCSB DB beyond metabolism and covering signaling pathways. Although metabolic and signaling networks are often investigated separately, based on our experience, incorporating different levels of biological data into one comprehensive and homogeneous source, and examining interplay occurring among them could lead to more realistic understanding of bioprocesses.

Bio DB unifies our current understanding of biochemical data, but it cannot explain the chemodiversity that we observe in living organisms. For example, plants synthesize a broad range of chemicals with various pharmacological applications. Even though scientists can isolate and measure these chemicals using analytical chemistry techniques, their synthesis pathways are often unknown. Not only the biosynthesis of chemicals is complicated, but also their degradation is a challenging task; even though many microorganisms are known to degrade these chemicals using unknown biochemical pathways. Therefore, we need to fill our biochemical knowledge gaps not only to understand natural processes, but also to unleash the potential of biochemistry for bioengineering purposes. Today, biochemical knowledge gap is beyond the scale to be addressed experimentally using biochemical assays. Therefore, it is essential to computationally propose

hypothesis and guide experimental efforts. We need computational methods that are able to learn from the available data and expand the boundaries of currently known biochemical space. The computational tool BNICE.ch, helps toward this aim by systematically exploring biochemistry and predicting novel metabolic reactions. BNICE.ch is a forerunner tool in predictive biochemistry which was developed more than fifteen years ago. Since then, its application and scope has been actively growing. The backbone of this algorithm is a collection of 850 expert-curated, generalized enzymatic reaction rules that mimic the action of enzymes and transform *in silico* substrates into products. Enzymatic reaction rules, a powerful network generation algorithm, and stored biochemical information in LCSB DB, make BNICE.ch an exceptional tool for filling gaps in biochemistry. In order to keep up with the latest biochemical discoveries, reaction rules and biochemical data in LCSB DB are regularly updated. Currently, we are at the stage that we can apply reaction rules on metabolites and (i) reconstruct all the processable known biochemical reactions, and (ii) predict novel metabolic reactions due to the promiscuity of reaction rules. In chapter 3, we focused on the application of BNICE.ch on different sets of compounds (such as biological molecules, drugs or chemical compounds), called ATLASx series of projects. In the first ATLAS database, reaction rules were applied on all the compounds in KEGG database, which resulted in reconstruction of around 5.2k known metabolic reactions, prediction of more than 130k novel biotransformations, and integration of more than 4k orphan compounds (compounds without any known biochemical activity). Over the years, the new biochemical discoveries and new reactions added to the databases such as KEGG have validated hundreds of novel ATLAS reactions. Moreover, a recent study by Yang, et.al [6] approved the activity of two novel ATLAS reactions in the context of one carbon assimilation pathways. Since the first publication of ATLAS in 2016 and up until today, more than 150 academic or industrial groups have requested to access it. The interest of research community on the first ATLAS encouraged us to publish an updated version in 2018. In the updated ATLAS, we applied the reaction rules on the last release of KEGG compounds, and we used an improved method for annotation of novel reactions with enzymes (BridgIT). The updated ATLAS incorporates 149k reactions and integrates 4.5k orphan compounds into network of metabolism. However, we quickly realized that using only KEGG compounds limits the type of problems we can address. For example, many drugs and chemicals are not listed in KEGG, and therefore they cannot be reached by ATLAS. Hence, we expanded the methodology of ATLAS to all biological and bioactive molecules (bioATALS) in LCSB DB and further to chemical compounds (chemATLAS). Basically, ATLAS projects attempt to use biochemical knowledge and biochemical reaction principles to map the hypothetical vicinity of known biochemical databases to address the vast amount of metabolic “dark matter”. We first predicted 1.6 million biochemically possible biotransformations between biological and bioactive compounds using bioATLAS, and then predicted more than 3.6 million reactions that involved compounds from the chemical compound space, resulting in a total amount of ~5.2 million reactions in chemATLAS. From this new wealth of information, we extracted insightful numbers on the reactivity and connectivity of biologically relevant molecules, and provide public access to our ATLASx database

(https://lcsb-databases.epfl.ch/Atlas2)." Moving forward, in the later stages of ATLASx projects, we recommend expanding our set of biochemical reaction rules to the known reaction mechanisms of organic chemistry. This will enable scientists to consider non-enzymatic and spontaneous reactions in ATLASx's reaction network. In addition using these upgraded networks, one can determine which steps in the synthesis pathway could be achieved by chemical reactions. The result will be a hybrid tool that smoothly connects biochemistry with chemistry.

The novel hypothetical metabolic reactions predicted by advanced computational tools (ATLAS reactions) allow expanding the known space of metabolism. However, the predicted novel reactions lack an associated enzyme for their catalysis, which limits their further application. Beside these novel biochemical reactions, half of the enzymatic reactions that are catalogued in the KEGG database remain orphan, i.e., that there is no known enzyme for their catalysis [7]. Such knowledge gaps limit the utility of the pathways that involve orphan reactions in synthetic biology and metabolic engineering. Furthermore, even if enzimens in the whole pathway are characterized, not all the enzymes are phylogenetically compatible to the others. This reality limits the implementation of even non-orphan reactions from different family of species. In chapter 4, we introduced a new reaction similarity method for assigning protein sequence to orphan and novel reactions, named BridgIT[7]. BridgIT uses reaction fingerprints to compare enzymatic reactions and is inspired by the "lock and key" principle that is used in protein docking methods; wherein the enzyme binding pocket is the "lock" and the ligand is a "key". If a molecule has the same reactive sites and a similar surrounding structure as the native substrate of a given enzyme, it is then rational to expect that the enzyme will catalyze the same biotransformation on this molecule. Following this logic, BridgIT uses the structural similarity of the reactive sites of participating substrates together with their surrounding structure as a metric for assessing the similarity of enzymatic reactions. This consept is substrate-reactive-site centric. Its reaction fingerprints reflect the specificities of biochemical reactions that arise from the type of enzymes catalyzing them. BridgIT introduces an additional level of specificity into reaction fingerprints by capturing critical information about the enzyme binding pocket. More precisely, BridgIT allows us to capture approximately the 2D structure of the enzyme binding pocket by incorporating the information about sequences of atoms and bonds around the substrate reactive site. The fact that BridgIT similarity calculations can be performed only by structural information of a reactions , is interesting and makes it a promising method for annotation of orphan reactions and filling the gaps of metabolic pathways.

A follow-up project to BridgIT, could involve the analysis of orphan protein sequences and evaluate their potential enzymatic activity, which is called BridgIT^{+} method here (

Figure 7.1). While the number of fully sequenced genomes are rapidly increasing, their functional annotation lags behind [8]. Approximately, the function of 30% to 50% of a normal genome is unknown[9]. Moreover, it has been estimated that 30% of unannotated sequences have metabolic function [9], indicating important

knowledge gaps in our understanding of cellular metabolism. However precise characterization of an uncharacterized protein requires extensive *in vitro* and *in vivo* experiments. Computational methods by significantly reducing time and cost of this process are very attractive approaches for protein characterization [10, p. 4].

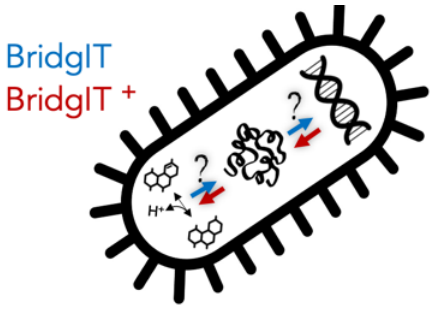


Figure 7.1: Conceptual comparison of BridgIT and BridgIT^{+} applications. BridgIT method annotates orphan reactions with protein sequences. Conversely, BridgIT^{+} method will aim to annotate orphan (or hypothetical) protein sequences with biochemical functions.

Computational approaches today are widely focused on inferring biochemical functionality from sequence or structural homology [11]. They assume if two sequences (or structures) are more similar than what is expected by chance, they could have evolved from a common ancestor [12], [13]. Nevertheless, following this logic, it is difficult to identify, (i) functional similarity of ortholog proteins (i.e. protein sequences with the same ancestor diverged as a result of speciation [11]), or (ii) difference in functionality of paralogs (i.e. homologue sequences diverged as a result of duplication [11]). Therefore, functional annotation only based on homology is not enough. We need better enzymatic descriptors to guide homology search for functional annotation of enzymatic sequences [11], [14]. Following this argument, PRIAM method [11] employs specific enzymatic profiles for protein sequence annotation. Enzymatic profiles distill the functional knowledge embedded in a group of sequences associated to one EC number. The trained enzymatic profiles for each EC number are later used for similarity evaluation and annotation of uncharacterized proteins. Nevertheless, the EC numbers are not the ideal criteria for this purpose. EC numbers are designed to systematically classify enzymatic activities based on type, functional group, involved cofactors and substrates. Thus, they don't capture the evolutionary changes of enzymes. This hypothesis is confirmed by the fact that approximately 40% of enzymes have evolved to completely new EC numbers (different in first digit of EC) [15]. Therefore, relation between chemistry and protein sequence of enzymes is more complicated than what is believed [15]. BridgIT method unbiasedly discovers the secondary functions of enzymes and quantifies their promiscuity [7]. We suggest, enriching enzymatic profiles with the knowledge of closest promiscuous enzymes using workflow in figure 7.2.

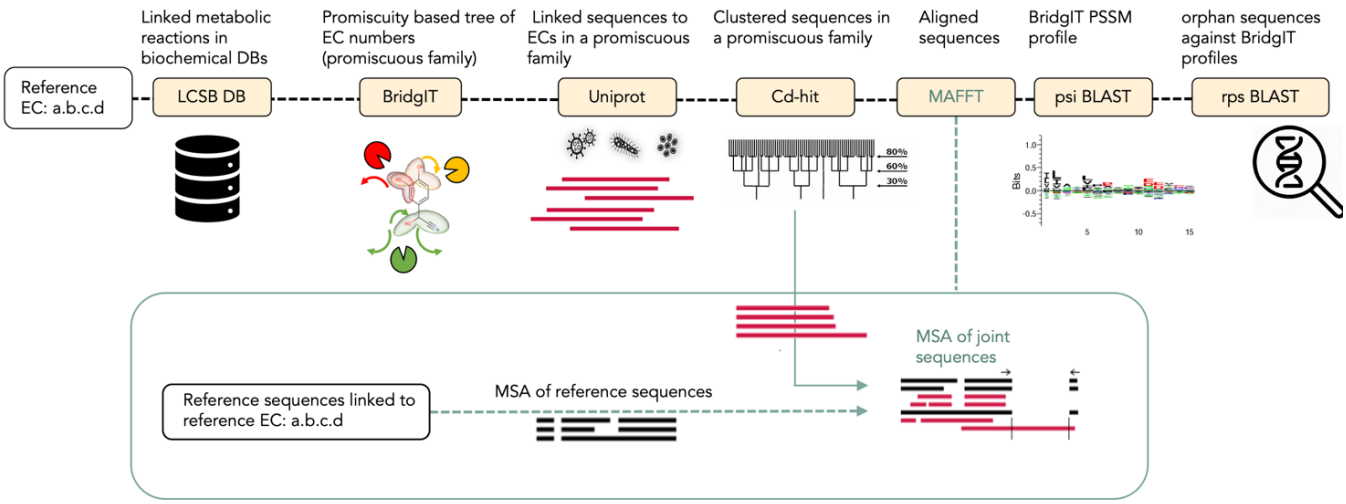


Figure 7.2: Suggested workflow for BridgIT+ method. The input of this workflow is an EC number (reference EC). The reference EC number is used to query LCSB DB in order to find all linked biochemical reactions. Next, BridgIT finds the most similar reactions to the extracted biochemical reactions using reactive site centric fingerprints. The EC numbers associated to the most similar reactions designate the candidate promiscuous activities. The ranked list of EC number will be used to collect sequences from protein databases (such as uniport [16]). Then, sequence clustering tools such as cd-hit [17] will be applied to group proposed promiscuous sequences into similar clusters. We suggest using MAFTT method [18] to align reference sequences with clustered promiscuous sequences. MAFTT method begins by aligning the reference sequences (MSA of reference sequences), then it aligns the cluster of promiscuous sequences to the reference MSA (joint MSA). Joint MSA preserves the biochemical knowledge of the reference EC number and on top of that takes into account promiscuity. Finally, Joint MSA is used for generation of enzymatic profiles (BridgIT profiles). After creation of BridgIT profiles for all EC numbers, they can be used for the annotation of whole genome using rps BLAST.

One of the main applications of BridgIT is in enzymatic annotation of metabolic reactions in bioproduction of valuable compounds. Fully annotated pathways have the capability to be implemented in organisms and optimized for bioproduction purposes. The bioproduction of added value compounds using cell factories could be an intriguing solution to few environmental related issues such as the effects of traditional chemical synthesis on global warming. In chapter 5, we took a step beyond the theoretical studies, and we tested the performance of BridgIT in practice to address challenges in bioproduction via two case studies: adipic acid and plant natural products (PNPs).

Adipic acid is a dicarboxylic acid used as a precursor for the production of nylon. About 2.5 billion kilograms of this compound is produced annually worldwide [19] by catalytic oxidation of benzene derivatives. bypeoducts of this method are nitrous oxide, which contributes to the greenhouse effect and ozone layer depletion [20]. Replacing the current petrochemical approach with bioproduction of adipic acid is a desirable alternative [21]. We used BridgIT method to annotate a new biosynthetic route towards adipic acid via the

lysine pathway in the yeasts *Saccharomyces cerevisiae* and *Yarrowia lipolytica*. *Y. lipolytica* was engineered by over-expression of homocitrate synthase *YALI0F31075g* (E.C.2.3.3.14), homoaconitate hydratase *YALI0E02728g* (E.C.4.2.1.36), di- and tri-carboxylic acids mitochondrial transporters *YALI0D02629g* and *YALI0F26323g*, and by expression of codon-optimized adipate-semialdehyde dehydrogenases from *Acinetobacter* sp. and *Pseudomonas* sp. The engineered strain produced up to 0.2 mg/L of adipic acid in mineral medium with glucose as the sole carbon source and up to 30 mg/L adipic acid in municipal solid waste hydrolyzate.

In the second case study, we developed a computational workflow to identify potential derivatives of intermediate compounds of a given biosynthetic pathway and subsequently predict enzyme candidates that may carry out the desired transformation(s). In contrast to previously reported retrobiosynthesis studies, in which a predicted pathway to a given target is generated, our workflow begins with a set of starting compounds (i.e., the intermediates of a heterologous biosynthetic pathway) and determines a set of novel target compounds and associated pathways that can be generated. The method expands the chemical space around a pathway of interest using BNICE.ch to create a map of all compounds accessible with biochemical reactions and then identifies enzymes capable of carrying out the desired transformations on the prioritized set of compounds using BridgIT. As an example, we applied this workflow to the reconstructed noscapine biosynthetic pathway in yeast. We narrowed our search to enzyme candidates capable of producing (S)-tetrahydropalmatine, a PNP found in plants of the genus Corydalis that has been shown to possess analgesic and anxiolytic effects and are known as a potential treatment for opiate addiction. After experimental evaluation of top BridgIT enzyme candidates in yeast strains, the two top predicted enzymes enabled production of (S)-tetrahydropalmatine. To the best of our knowledge, our work describes the first use of a computational workflow to expand a heterologous biosynthetic pathway to produce additional compounds.

The findings stemming from these two case studies show the value of chem-informatic tools in design stage of a design-build-test-learn cycle in engineering biology. The described pipeline can be used for systematic exploration of alternatives in production of many chemically complex compounds spanning diverse therapeutic activities. For any target compound, computational design workflow consists of (i) discovery of pathways, (ii) predicting enzymes and (iii) evaluating feasibility of proposed pathway. In future, advancement of technology and development of our knowledge could lead to further optimization in each step of this workflow. Particularly, enhancement of computational power of pathway prediction tools to explore beyond linear pathways and ability to analyse branched and complex pathways would open new doors to an untapped space of metabolism. Moreover, to further optimize the activity of proposed enzyme candidates, we suggest coupling the results of BridgIT method with the protein design tools (such as Rosetta Design software, YASARA and FoldX [22]–[24]). The mentioned tools use the 3D structure of proteins and by performing free energy state calculations and molecular dynamic simulations, they are able to predict the

outcome of amino acid substitutions on protein structures. Their outcome offers a list of candidate variants for identification of improved enzymes. The proposed enzymes for catalyzation of orphan reactions by BridgIT, can serve as good initial sequences for further enzyme engineering using protein design tools.

Understanding metabolism at the molecular level is one of the most important element in drug discovery and drug development. In chapter 6 of this thesis, we focused on application of the predictive biochemistry tools in drug design. Discovering new non-toxic drugs is essential to treat diseases and infections, target drug resistance, and develop personalized treatments. However, identifying, testing, and approving a single small molecule can take decades and billions of dollars—and there is still a high risk that the proposed drug candidate fails. There is an urgent need to define strategies that accelerate the discovery of new, safe, and effective drugs. The computational screening of *all* possible targets and molecules can help toward this aim. Most computational approaches to date have focused on molecular structures without considering the reactivity of the molecules in a cell. However, reactive site information and drug metabolism determine which enzymes the drugs will target, the drug's metabolic fate or degradation, and the potential source of its toxicity and side effects. Understanding drug effects in the context of cellular metabolism also offers great promise in evaluating the reactivity of a new small molecule, the druggability of an enzyme, and the possibility of drug repurposing. Yet, the *in silico* mechanistic analysis of drug biochemistry is relatively unexplored, and no major large-scale computational studies of drug metabolism in cells have ever been performed. To systematically illuminate the metabolism and *all* enzymatic targets of known drugs and hypothetical prodrugs, we have performed the first large-scale computational analysis of drug biochemistry and toxicity evaluation in the context of human metabolism. To this end, we employed proven tools for analyzing the neighboring atoms around enzyme reactive sites (BridgIT and BNICE.ch). The analysis involved over 250,000 small molecules, and was a major technical effort spanning the curation and computation of bio- and physico-chemical drug properties. We assembled this in an open-source drug resource, NICEdrug.ch, that can generate drug metabolic reports and can be easily accessed and used by researchers, clinicians, and industrial partners around the world. Excitingly, NICEdrug.ch revealed for the first time that known drugs, such as anticancers, cholesterol reducing drugs, antimalarials, and drugs against COVID-19, share millions of reactive sites with native human metabolites and can hence act as competitive inhibitors. As a first case study, NICEdrug.ch showed unexplored sources of toxicity for one of the most used anticancer drugs *5-FU* and suggested an alternative treatment to alleviate its toxic side-effects. Secondly, we studied the reactivity of statins, which are broadly used as cholesterol reducing drugs, and showed that comparing the reactive sites can accurately cluster two families of statins that are similar in overall molecular structure but have different reactivities. As a third proof-of-principle demonstration, we applied NICEdrug.ch to identify new drug candidates for targeting liver-stage malaria parasites and minimize their impact on the human host cell. To our surprise, the top identified drug candidate by NICEdrug.ch has already been used for *Escherichia* and *Streptococcus* infections without appreciable side effects. In our fourth demonstration, NICEdrug.ch

identified over 1,300 drugs and molecules in food substances to fight COVID-19 and explained their inhibitory mechanism. Among our results we have both experimentally validated drugs, like statins, and new ones, like N-acetylcysteine. NICEdrug.ch currently includes metabolic information for human cells, *Plasmodium*, and *Escherichia coli*, and it is easily expandable to other organisms in the future. NICEdrug.ch also allows for the input of new small molecules to tailor it for a user's needs. This major informational database will be updated regularly to provide the latest information on drugs and drug metabolism.

We close this thesis by highlighting the fact that computational mindset brings a new order into our understanding of life, allows us to put together the individual insights, create a reference map and see the big picture [25]. During this PhD work, we contributed to the field of computational biology through developing new methods and databases for enhancing our understanding of metabolism. Utility of these tools provides an unprecedented potential for large-scale and systematic analysis of metabolism.

7.1 References

[1] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "KEGG as a reference resource for gene and protein annotation," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D457–D462, Jan. 2016, doi: 10.1093/nar/gkv1070.

[2] R. Caspi *et al.*, "The MetaCyc database of metabolic pathways and enzymes," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D633–D639, Jan. 2018, doi: 10.1093/nar/gkx935.

[3] L. Jeske, S. Placzek, I. Schomburg, A. Chang, and D. Schomburg, "BRENDA in 2019: a European ELIXIR core data resource," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D542–D549, Jan. 2019, doi: 10.1093/nar/gky1048.

[4] D. S. Wishart *et al.*, "HMDB 4.0: the human metabolome database for 2018," *Nucleic Acids Res.*, vol. 46, no. Database issue, pp. D608–D617, Jan. 2018, doi: 10.1093/nar/gkx1089.

[5] K. Munir and M. Sheraz Anjum, "The use of ontologies for effective knowledge modelling and information retrieval," *Appl. Comput. Inform.*, vol. 14, no. 2, pp. 116–126, Jul. 2018, doi: 10.1016/j.aci.2017.07.003.

[6] X. Yang *et al.*, "Systematic design and in vitro validation of novel one-carbon assimilation pathways," *Metab. Eng.*, vol. 56, pp. 142–153, Dec. 2019, doi: 10.1016/j.ymben.2019.09.001.

[7] N. Hadadi, H. MohammadiPeyhani, L. Miskovic, M. Seijo, and V. Hatzimanikatis, "Enzyme annotation for orphan and novel reactions using knowledge of substrate reactive sites," *Proc. Natl. Acad. Sci.*, p. 201818877, Mar. 2019, doi: 10.1073/pnas.1818877116.

[8] R. S. Baric, S. Crosson, B. Damania, S. I. Miller, and E. J. Rubin, “Next-Generation High-Throughput Functional Annotation of Microbial Genomes,” *mBio*, vol. 7, no. 5, Nov. 2016, doi: 10.1128/mBio.01245-16.

[9] M. Griesemer, J. A. Kimbrel, C. E. Zhou, A. Navid, and P. D’haeseleer, “Combining multiple functional annotation tools increases coverage of metabolic annotation,” *BMC Genomics*, vol. 19, no. 1, p. 948, Dec. 2018, doi: 10.1186/s12864-018-5221-9.

[10] Md. S. Ahmed, Md. Shahjaman, E. Kabir, and Md. Kamruzzaman, “Structure modeling to function prediction of Uncharacterized Human Protein C15orf41,” *Bioinformation*, vol. 14, no. 5, pp. 206–212, May 2018, doi: 10.6026/97320630014206.

[11] C. Claudel-Renard, C. Chevalet, T. Faraut, and D. Kahn, “Enzyme-specific profiles for genome annotation: PRIAM,” *Nucleic Acids Res.*, vol. 31, no. 22, pp. 6633–6639, Nov. 2003, doi: 10.1093/nar/gkg847.

[12] W. R. Pearson, “An Introduction to Sequence Similarity (‘Homology’) Searching,” *Curr. Protoc. Bioinforma. Ed. Board Andreas Baxevanis Al*, vol. 0 3, Jun. 2013, doi: 10.1002/0471250953.bi0301s42.

[13] V. Sangar, D. J. Blankenberg, N. Altman, and A. M. Lesk, “Quantitative sequence-function relationships in proteins based on gene ontology,” *BMC Bioinformatics*, vol. 8, no. 1, p. 294, Aug. 2007, doi: 10.1186/1471-2105-8-294.

[14] C. Bannert, A. Welfle, C. aus dem Spring, and D. Schomburg, “BrEPS: a flexible and automatic protocol to compute enzyme-specific sequence profiles for functional annotation,” *BMC Bioinformatics*, vol. 11, no. 1, p. 589, Dec. 2010, doi: 10.1186/1471-2105-11-589.

[15] S. Martínez Cuesta, S. A. Rahman, N. Furnham, and J. M. Thornton, “The Classification and Evolution of Enzyme Function,” *Biophys. J.*, vol. 109, no. 6, pp. 1082–1086, Sep. 2015, doi: 10.1016/j.bpj.2015.04.020.

[16] S. Pundir, M. Magrane, M. J. Martin, C. O’Donovan, and The UniProt Consortium, “Searching and Navigating UniProt Databases: Searching and Navigating UniProt Databases,” in *Current Protocols in Bioinformatics*, A. Bateman, W. R. Pearson, L. D. Stein, G. D. Stormo, and J. R. Yates, Eds. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2015, p. 1.27.1-1.27.10.

[17] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, “CD-HIT: accelerated for clustering the next-generation sequencing data,” *Bioinforma. Oxf. Engl.*, vol. 28, no. 23, pp. 3150–3152, Dec. 2012, doi: 10.1093/bioinformatics/bts565.

[18] K. Katoh, J. Rozewicki, and K. D. Yamada, “MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization,” *Brief. Bioinform.*, vol. 20, no. 4, pp. 1160–1166, Jul. 2019, doi: 10.1093/bib/bbx108.

[19] K. Raj, S. Partow, K. Correia, A. N. Khusnutdinova, A. F. Yakunin, and R. Mahadevan, “Biocatalytic production of adipic acid from glucose using engineered Saccharomyces cerevisiae,” *Metab. Eng. Commun.*, vol. 6, pp. 28–32, Jun. 2018, doi: 10.1016/j.meteno.2018.02.001.

[20] R. W. Portmann, J. S. Daniel, and A. R. Ravishankara, “Stratospheric ozone depletion due to nitrous oxide: influences of other gases,” *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 367, no. 1593, pp. 1256–1264, May 2012, doi: 10.1098/rstb.2011.0377.

[21] J.-L. Yu, X.-X. Xia, J.-J. Zhong, and Z.-G. Qian, “Direct biosynthesis of adipic acid from a synthetic pathway in recombinant Escherichia coli,” *Biotechnol. Bioeng.*, vol. 111, no. 12, pp. 2580–2586, 2014, doi: 10.1002/bit.25293.

[22] R. Das and D. Baker, “Macromolecular modeling with rosetta,” *Annu. Rev. Biochem.*, vol. 77, pp. 363–382, 2008, doi: 10.1146/annurev.biochem.77.062906.171838.

[23] F. Richter, A. Leaver-Fay, S. D. Khare, S. Bjelic, and D. Baker, “De Novo Enzyme Design Using Rosetta3,” *PLoS ONE*, vol. 6, no. 5, p. e19230, May 2011, doi: 10.1371/journal.pone.0019230.

[24] E. Krieger, G. Koraimann, and G. Vriend, “Increasing the precision of comparative models with YASARA NOVA--a self-parameterizing force field,” *Proteins*, vol. 47, no. 3, pp. 393–402, May 2002.

[25] F. Markowetz, “All biology is computational biology,” *PLOS Biol.*, vol. 15, no. 3, p. e2002050, Mar. 2017, doi: 10.1371/journal.pbio.2002050.

Chapter 8 **Appendix**

Supplementary Tables

Table 8.1: Quality of reactions in different sources based on mass balance and EC annotation.

Database	# reactions with EC assigned	# Balanced reactions (% Reconstructed with BNICE.ch enzymatic rules)
HMR	3,417	3,177 (40.2%)
MetaCyc	9,614	7,879 (78.4%)
KEGG	9,667	9,010 (80.2%)
MetaNetX	14,194	12,733 (54.5%)
Reactome	342	406 (52.4%)
Rhea	10,808	10,401 (58.1%)
Model SEED	9,816	14,290 (47.4%)
BKMS	15,493	10,556 (73.3%)
BiGG models	3,874	3,445 (27.5%)
Brenda	6,629	6,825 (63.1%)
Total	27,107	25,296 (46.5%)

Table 8.2, 8.3: Comparison of EC predictor tools for two benchmark reactions

The tables 8.2 and 8.3 show the performance of different methods (BridgIT, EC-BLAST, Selenzyme and E-zyme2) for two main challenges that are represented by the class of reactions:

- Reactions with a similar structure on the substrate and product side.
- Multi-substrate multi-product reactions (a subset of reactions with more than one substrate)

Note that these two classes of reactions are ubiquitous in biochemical networks. For comparisons between methods, we took an example reaction of each class, i.e., R00722 (2.7.4.6) for the first class and R07500 (2.5.1.115) for the second class. For the two benchmark reactions, we ranked the similar reactions proposed by each of methods according to the corresponding similarity scores, and top 100 similar reactions proposed by each method were used for comparisons. We used the following criteria to quantitatively compare these tools:

1. The number of matched 4^{th} level EC numbers between the benchmark reaction and reactions proposed by the tested method. We introduced this criterion because the reactions that share the same 4^{th} level EC number, in most cases have a similar mechanism, cofactors and the structure of substrates.
2. The number of matched 3^{rd} level EC numbers between the benchmark reaction and the reactions proposed by the tested method - the reactions that share the same 3^{rd} level EC number, in most cases have a similar mechanism and cofactors, but less structural similarity of substrates compared to the 4^{th} level matched EC numbers.
3. The number of unique 4^{th} level EC numbers in the set of reactions proposed by the tested method that had matched 3^{rd} level EC numbers in Criterion 2. The higher ratio between this number and the number from criterion 3, the method has a wider scope of predicted enzymes. The maximal value of this ratio is 1.
4. A Receiver Operating Characteristics (ROC) and the Area Under the Curve (AUC). Each of the compared methods predicted for the benchmark reaction a set of similar reactions together with their similarity scores. Using this information, we constructed the ROC curves and computed AUC for each of the methods. Therein, following the approach proposed in the manuscript on the EC Blast method, we considered a result as true positive if a predicted reaction by the tested method

matched the 3rd level EC number of the benchmark reaction. The obtained ROC curves allow us to assess the robustness and confidence levels of each tool's predictions.

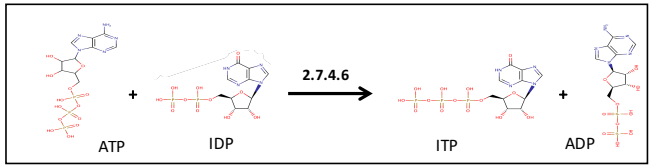
5. Mapping each input reaction to itself (whenever the input reaction is not orphan) – as a basic functionality of enzyme annotation methods.

Table 8.2: Comparison of EC predictor tools for benchmark reaction 1 exemplifying the first class of reactions characterized by a very similar structure of substrates and products.

R00722

2.7.4.6

Challenge: similar structure in product and substrate side



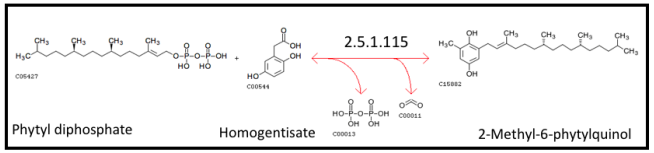
Existing tools		Description	Self recognized	# similar rxns		# similar unique 4 level EC with the same 3rd level	score		Best ECs
				4 level EC	3 level EC		Best	min	
BridgIT		Based on BridgIT Fingerprint which includes information of reactive site and its neighborhood.	yes	11	35	22	1	0.51	2.7.4.6
EC BLAST	Bond changes (BC)	Based on bond change similarity:bond formed/cleaved, bond order change and stereo change	yes	11	15	10	1	1	2.7.1.73 2.7.4.6 2.7.4.3 3.6.1.60 (22 more)
	Reaction center (RC)	Based on reaction center similarity. Reaction centers are connected to a bond that is broken/formed or the order of bond or its stereo is changed.	yes	11	30	19	1	0.77	2.7.4.6 2.7.4.18 2.7.4.15 2.7.6.2
	Both BC and RC	Based on both bond change and reaction center methods.	yes	11	30	19	1	0.88	2.7.4.6 2.7.4.18. 2.7.4.15. 2.7.6.2
	Structural similarity	Based on substructure similarity. All the molecules in the query reaction are compared to all those in a target.	yes	7	10	8	0.94	0.78	2.7.1.73
Selenzyme	Rdkit	based on Rdkit fingerprints	yes	0	3	3	0.92	0.88	2.7.1.73
	Morgan	based on pattern fingerprints	yes	0	8	4	0.91	0.85	2.7.1.25
	Pattern	based on Morgan fingerprints	yes	0	9	8	0.7	0.63	2.7.4.8
E-zyme2		Based on structures of substrate-product pair (reactant pair).	-	-	-	-	-	-	-

Table 8.3: Comparison of EC predictor tools for benchmark reaction 2 exemplifying the class of multi-substrate multi-product reactions.

R07500

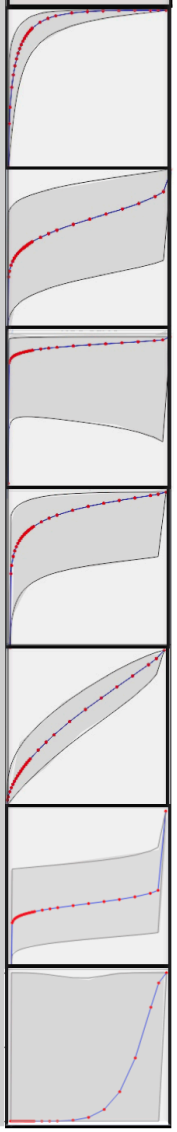
2.5.1.115

Challenge: Bi-substrates



Existing tools		Description	Self recognized	# similar rxns		# similar unique 4 level EC with the same 3rd level	score		Best ECs	AUC	95% CI*
				4 level EC	3 level EC		Best	min			
BridgIT		Based on BridgIT Fingerprint which encompass information of reactive site and its neighborhood.	yes	0	39	26	0.88	0.1	2.5.1.117 2.5.1.116	0.95	0.86-0.96
EC BLAST	Bond changes (BC)	Based on bond change similarity: bond formed/cleaved, bond order change and stereo change.	yes	0	6	6	1	0.79	2.5.1.117 2.5.1.116	0.68	0.33-0.91
	Reaction center (RC)	Based on reaction center similarity. Reaction centers are connected to a bond that is broken/formed or the order of bond or its stereo is changed.	yes	0	5	5	1	0.56	2.5.1.117 2.5.1.116	0.94	0.41-0.99
	Both BC and RC	Based on both bond change and reaction center methods.	yes	0	7	7	1	0.59	2.5.1.117 2.5.1.116	0.87	0.5-0.97
	Structural similarity	Based on substructure similarity. All the molecules in the reaction are compared to all those in a target.	yes	0	45	36	0.69	0.39	2.5.1.117 2.5.1.116	0.59	0.46-0.71
Selenzyme		Reaction similarity is based on 3 types of Rdkit fingerprints, called: Rdkit, Pattern and Morgan	yes	0	10	7	0.94	0.5	2.5.1.117 2.5.1.116	0.40 ***	0.18-0.68 (***)
E-zyme2		Based on structures of substrate-product pair (reactant pair).	yes	0	9	4	0.61	0.33	2.5.1.39 2.5.1.93	0.2 ***	0.02-0.98 (***)

ROC curve with 95% confidence interval of the fitted ROC curve**



* 95% confidence interval of the fitted ROC curve.

** Estimated binormal roc curve, with lower and upper bounds of the asymmetric 95% confidence interval for true-positive fraction at a variety of false-positive fractions

*** ROC curve analysis for Selenzyne and E-zyme2 is not reliable due to a reduced set of studied reactions (Selenzyme proposed only 15 and E-zyme 18 similar reactions to R07500, whereas BridgIT and EC-Blast proposed more than 100 reactions).

The following tables are published on the Zenodo platform and accessible at https://doi.org/10.5281/zenodo.4004191).

Table 8.4: Strains used in adipic acid bioproduction chapter 5.2.

Table 8.5: Plasmids used in adipic acid bioproduction chapter 5.2.

Table 8.6: Biobricks used in adipic acid bioproduction chapter 5.2.

Table 8.7:Primers used in adipic acid bioproduction chapter 5.2.

Table 8.8:Heterologous genes discussed in adipic acid bioproduction chapter 5.2.

Table 8.9: The result of BLASTp algorithm with default settings to align the amino-acid sequence of the large and small subunits of all 121 annotated in KEGG metanogen homoaconitases to *Y. lypolytica* genome (CLIB122), related to chapter 5.2.

Table 8.10:Yeast strains used in *(S)*-tetrahydropalmatine bioproduction in chapter 5.3.

Table 8.11:Oligonucleotides used in *(S)*-tetrahydropalmatine bioproduction in chapter 5.3.

Table 8.12:Genes used in *(S)*-tetrahydropalmatine bioproduction in chapter 5.3.

Table 8.13:Plasmids used in *(S)*-tetrahydropalmatine bioproduction in chapter 5.3.

Table 8.14: LC-MS/MS multiple reaction monitoring (MRM) transitions and parameters used in *(S)*-tetrahydropalmatine bioproduction in chapter 5.3.

Table 8.15: sequences of codon-optimized genes used in *(S)*-tetrahydropalmatine bioproduction in chapter 5.3.

Table 8.16: Biochemical network generated by BNICE.ch – COMPOUNDS related to chapter 5.3.

Table 8.17: Biochemical network generated by BNICE.ch – REACTIONS related to chapter 5.3.

Table 8.18: Overview on network statistics related to chapter 5.3.

Table 8.19: Popularity analysis for all BIA compounds in the network, related to chapter 5.3.

Table 8.20:50 most popular compounds in the generated network, related to chapter 5.3.

Table 8.21:Additional information for 15 candidates targets one reaction step away from the initial pathway, related to chapter 5.3.

Table 8.22:(A) List of cofactors, (B) list of metabolites, and (C) list of E.C. numbers considered in BNICE.ch for the generation of reactions in the analysis of drug metabolism in a human cell, related to chapter 6.

Table 8.23: Metabolic neighborhood of 5-FU. (1) List of compounds in the 5-FU metabolic neighborhood including up to four reactions or steps away. (2) Description of reactions in the 5-FU metabolic neighborhood including up to four reactions or steps away, related to chapter 6.

Table 8.24: NICEdrug score between all molecules with reactive site of statins in NICEdrug.ch. Matrix of NICEdrug score between each pair of the whole set of 254 molecules in NICEdrug.ch with reactive site of statins, related to chapter 6.

Table 8.25: Description of nine drugs candidates for repurposing to replace statins based on NICEdrug.ch, related to Figure 6.5. These drugs can act as competitive inhibitors of HMG-CoA reductase, like statins, related to chapter 6.

Molecule name	**Molecule NICEdrug ID**	**Molecule ChEMBL ID**	**Molecule DrugBank ID**	**Molecule KEGG ID**	**Molecule SMILES**	**Most similar statin**	**NS[1]**
E-64	33476		DB04276, EXPT01317	C01341	CC(CC(C(=O)NCCCC[NH+]=C(N)N)NC(=O)C(CC(=O)O)O)C	Pravastatin	0.23
Nanaomycin D	3904013	ChEMBL MALARIA, CHEMBL 1988648	DB01668, EXPT02375	D04648	O=C1CC2C(O1)C1=C(C(O2)C)C(=O)c2c(C1=O)cccc2O	Pravastatin	0.27
2-Tridecanoyloxy-Pentadecanoic A	5459982		DB01814, EXPT02033		CCCCCCCCCCCCCC(CC(=O)O)OC(=O)CCCCCCCCCCCC	Pravastatin	0.6
OBP-801	96065750	CHEMBL 3126832	DB12279		O=C1OC2C=CCCSSCC(C(=O)NC(C(C1)O)C(C)C)NC(=O)C(NC(=O)C2)C	Pravastatin	0.36
Carnitine	1467871455	CHEMBL 1149, CHEMBL 1229656, CHEMBL 1620698, CHEMBL 172513, CHEMBL 503189	DB02648, EXPT00038	C00318,C00487,C15025	OC(C[N+](C)(C)C)CC(=O)O	Pravastatin	0.31
O-Acetylcarnitine	1467889778	CHEMBL 1358846, CHEMBL 1625375, CHEMBL 1697733	DB08842	C02571	OC(=O)CC(C[N+](C)(C)C)OC(=O)C	Pravastatin	0.47

Josamycin	1467981653	ChEMBL_NTD, CHEMBL 1326015, CHEMBL 1671903, CHEMBL 1995525, CHEMBL 224436, CHEMBL 2361089, CHEMBL 329011	DB01321	C12662,D01235	O=CCC1CC(C)C(O)C=CC=CCC(OC(=O)CC(C(C1OC1OC(C)C(C(C1O)N(C)C)OC1OC(C)C(C(C1)(C)O)OC(=O)CC(C)C)OC)OC(=O)C)C	Lovastatin acid	0.41
Plitidepsin	1468014507	CHEMBL 1773899, CHEMBL 451930	DB04977	C16862,D11032	CCC(C1NC(=O)C(NC(=O)C(N(C(=O)C2CCCN2C(=O)C(=O)C)C)CC(C)C)C(C)OC(=O)C(Cc2ccc(cc2)OC)N(C)C(=O)C2CCCN2C(=O)C(NC(=O)C(C(=O)C(OC(=O)CC1O)C(C)C)C)CC(C)C)C	Lovastatin acid	0.39

[1]NS: NICEdrug score between the candidate drug (column 1) and the most similar statin (column 7). This score considers the reactive site and its neighbourhood including up to seven atoms away (STAR Methods).

The following tables are published on the Zenodo platform and accessible at https://doi.org/10.5281/zenodo.4004191).

Table 8.26: Essential genes or enzymes and linked metabolites in liver-stage Plasmodium and a human cell. (A) List of essential genes and associated reactions in liver-stage Plasmodium, as obtained from the study (Stanway et al., 2019) (B) List of essential genes and associated reactions in a human cell, as obtained from the study (Wang et al., 2015) (C) List of metabolites linked to essential genes in liver-stage Plasmodium. (D) List of metabolites linked to essential genes in a human cell, related to chapter 6.

Table 8.27: Description of drugs, prodrugs, metabolites and enzymes analyzed in the study of malaria. (A) NICEdrug druggability analysis of essential genes or enzymes in liver-stage Plasmodium: all drugs sharing reactive-site centric similarity with the Plasmodium metabolites and comparison with human metabolites. (B) NICEdrug druggability analysis of essential genes or enzymes in liver-stage Plasmodium: all prodrugs (up to three steps away of 346 drugs) sharing reactive-site centric similarity with the Plasmodium metabolites and comparison with human metabolites. (C) Description of drugs and prodrugs identified in the malaria analysis with NICEdrug.ch and validated in the study by (Antonova-Koch et al., 2018) along with their similar Plasmodium metabolite and human metabolite, related to chapter 6.

Table 8.28:Hijacked human enzymes by SARS-CoV-2, and drugs and food-based compounds that can inhibit them based on the NICEdrug score. (A) Hijacked human proteins by SARS-CoV-2 as identified by (Gordon et al., 2020) with an annotated enzymatic function (E.C. number), also called here "SARS-CoV-2 hijacked enzymes". (B) NICEdrug druggability report for SARS-CoV-2 hijacked enzymes including all NICEdrug small molecules. (C) Best candidate drugs against COVID-19: NICEdrug druggability report for SARS-CoV-2 hijacked enzymes including drugs with NICEdrug score above 0.5 compared to the native human substrate. (D) Summary of NICEdrug best candidate drugs against COVID-19 and their classification according to the drug category in the KEGG database. (E) NICEdrug druggability report of SARS-CoV-2 hijacked enzymes including prodrugs (up to three steps away of any NICEdrug small molecule) with NICEdrug score above 0.5 compared to the native human substrate. (F) Best candidate food-based molecules against COVID-19: NICEdrug druggability report of SARS-CoV-2 hijacked enzymes including food-based molecules with NICEdrug score above 0.5 compared to the native human substrate. (G) Summary of the NICEdrug best candidate food-based molecules against COVID-19 and their classification according to the fooDB source, related to chapter 6.

Table 8.29:NICEdrug analysis of inhibitory mechanisms of currently used anti SARS-CoV-2 drugs. (A) All drug molecules and (B) prodrugs in NICEdrug.ch sharing reactive site with the native substrates of the human enzyme HDAC2 and their NICEdrug score with this substrate. (C) All molecules cataloged in fooDB sharing reactive site with the native substrates of the human enzyme HDAC2 and their NICEdrug score with this substrate. (D) All drug molecules and (E) prodrug molecules in NICEdrug.ch sharing reactive site with the native substrates of the human enzyme ACE2 and their NICEdrug score with this substrate. (F) All molecules cataloged in fooDB

sharing reactive site with the native substrates of the human enzyme ACE2 and their NICEdrug score with this substrate. (G) All molecules in NICEdrug.ch or cataloged in fooDB sharing reactive site with the native substrates of the human enzyme DNA-directed RNA polymerase and their NICEdrug score with this substrate, related to chapter 6.

Supplementary Figures

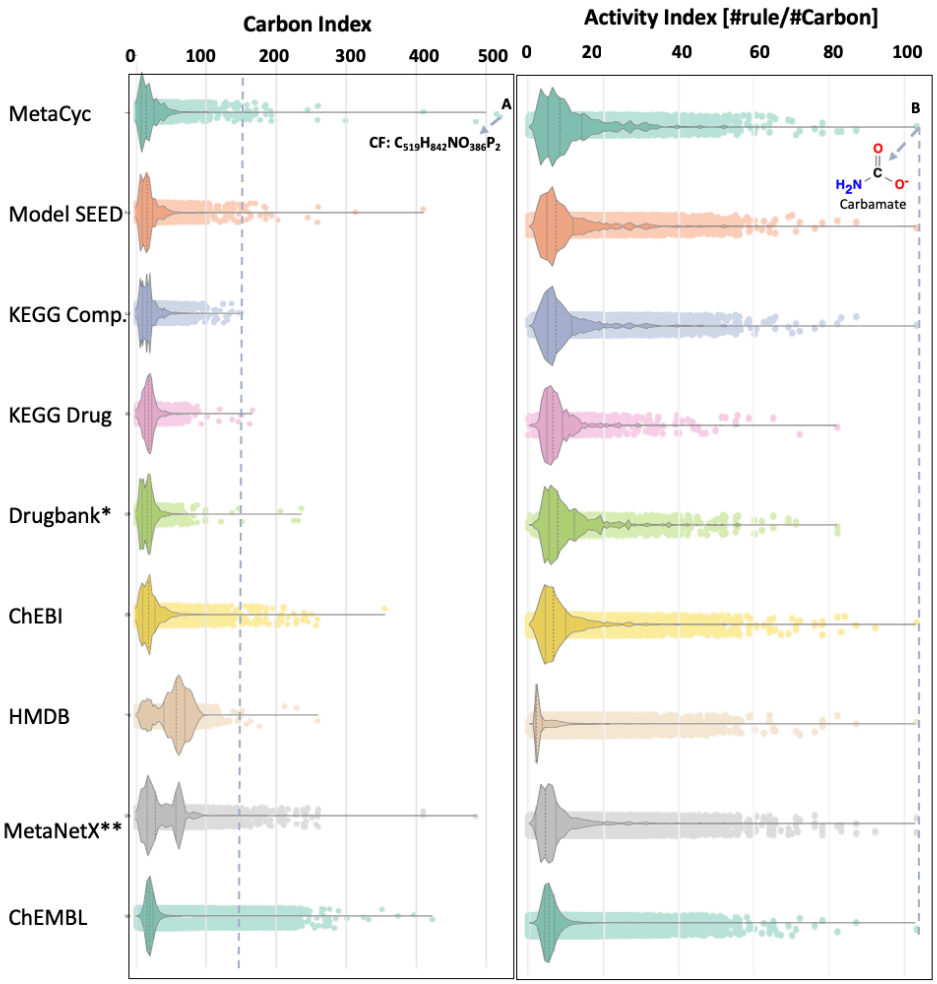


Figure 8.1: Overview on Compound databases in terms of number of carbons and activity, related to chapter 2.3.

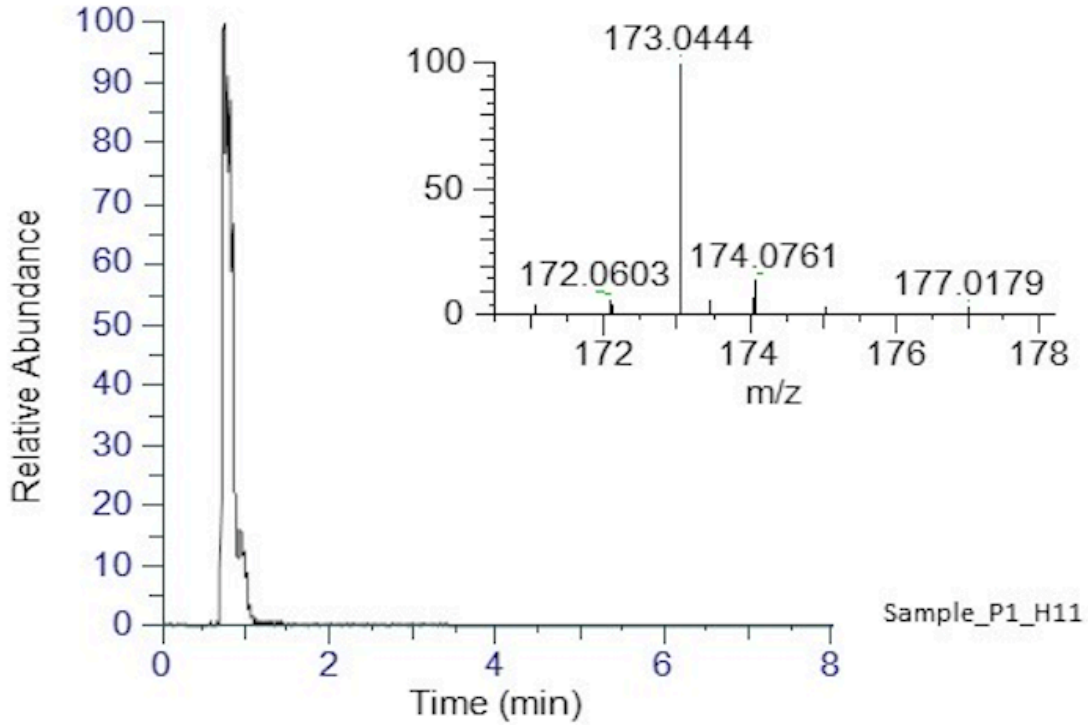
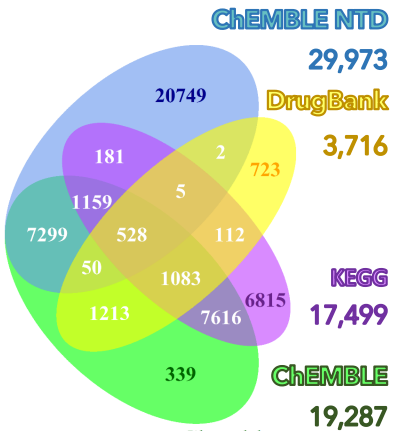


Figure 8.2: Extracted ion chromatogram of the 173.0444 m/z ion (mass of 2-oxopimelate, C7H8O5) in the sample in positive ionization mode, related to chapter 5.2.

A



B

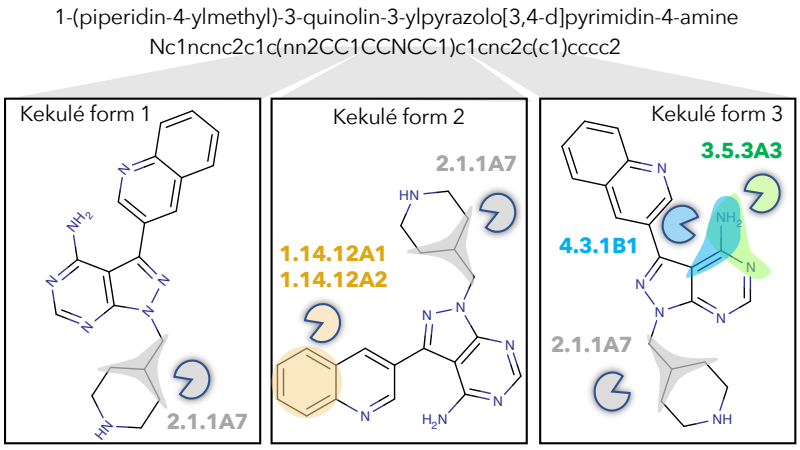


Figure 8.3: Overview of number of molecules in NICEdrug.ch and their structural curation. (A) Venn diagram showing the number of compounds in NICEdrug.ch and their source database: KEGG, DrugBank, ChEMBLE NTD, and ChEMBLE. (B) Representation on how different kekulé forms affect the identification of reactive sites and prediction of biological activity for an example molecule, related to chapter 6.

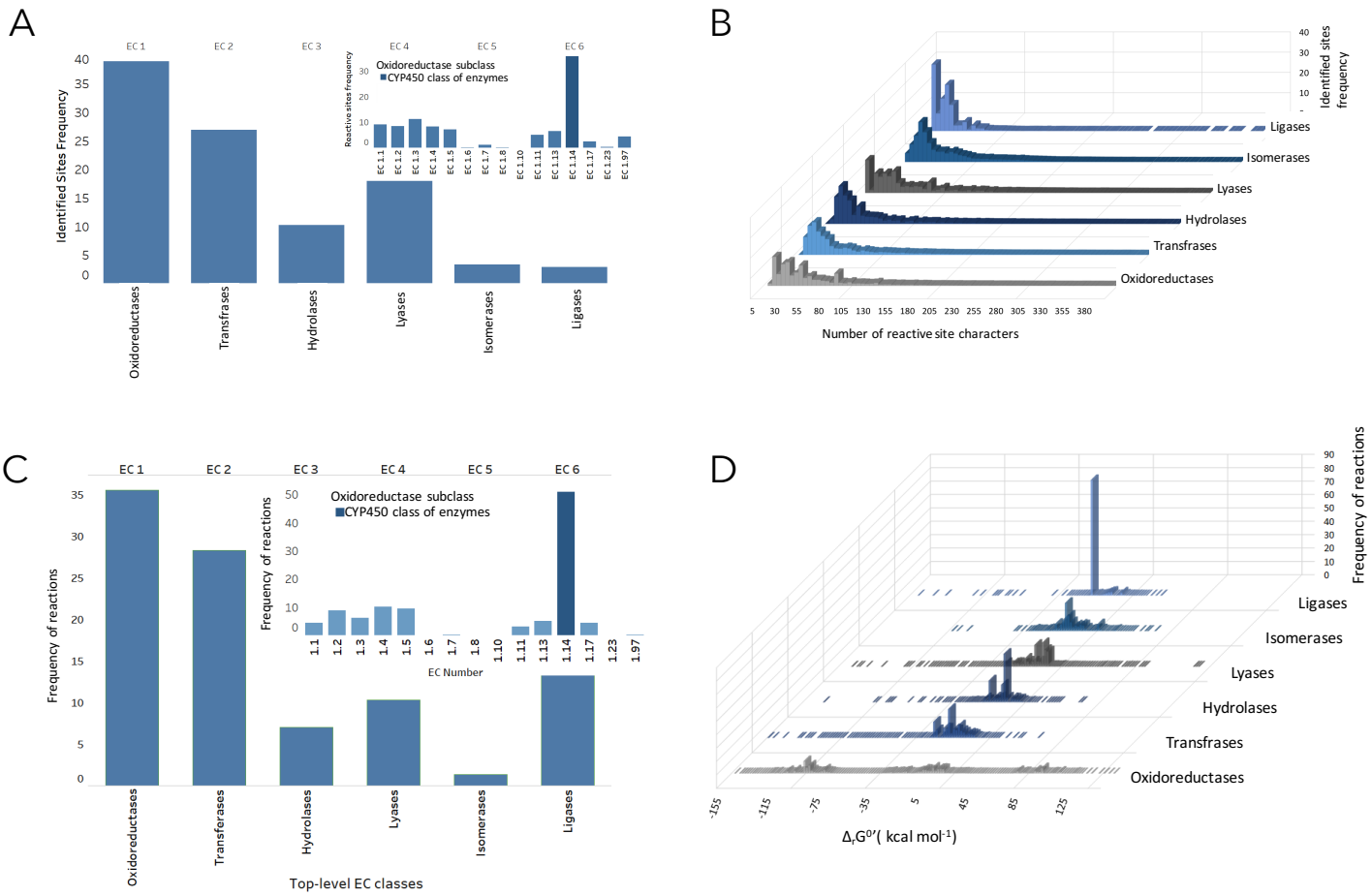


Figure 8.4: Distribution of reactive sites and metabolic reactions as of E.C. numbers linked to all molecules in NICEdrug.ch. (A) Distribution of reactive sites identified in all molecules of NICEdrug.ch among classes of E.C. numbers. (B) Specificity of reactive sites identified in drugs based on length and types of participating atoms. (C) Distribution of drug metabolic reactions based on class of E.C. number. (D) Distribution of Gibbs free energy for the drug metabolic reactions, which are the reactions linked to all molecules of NICEdrug.ch, related to chapter 6.

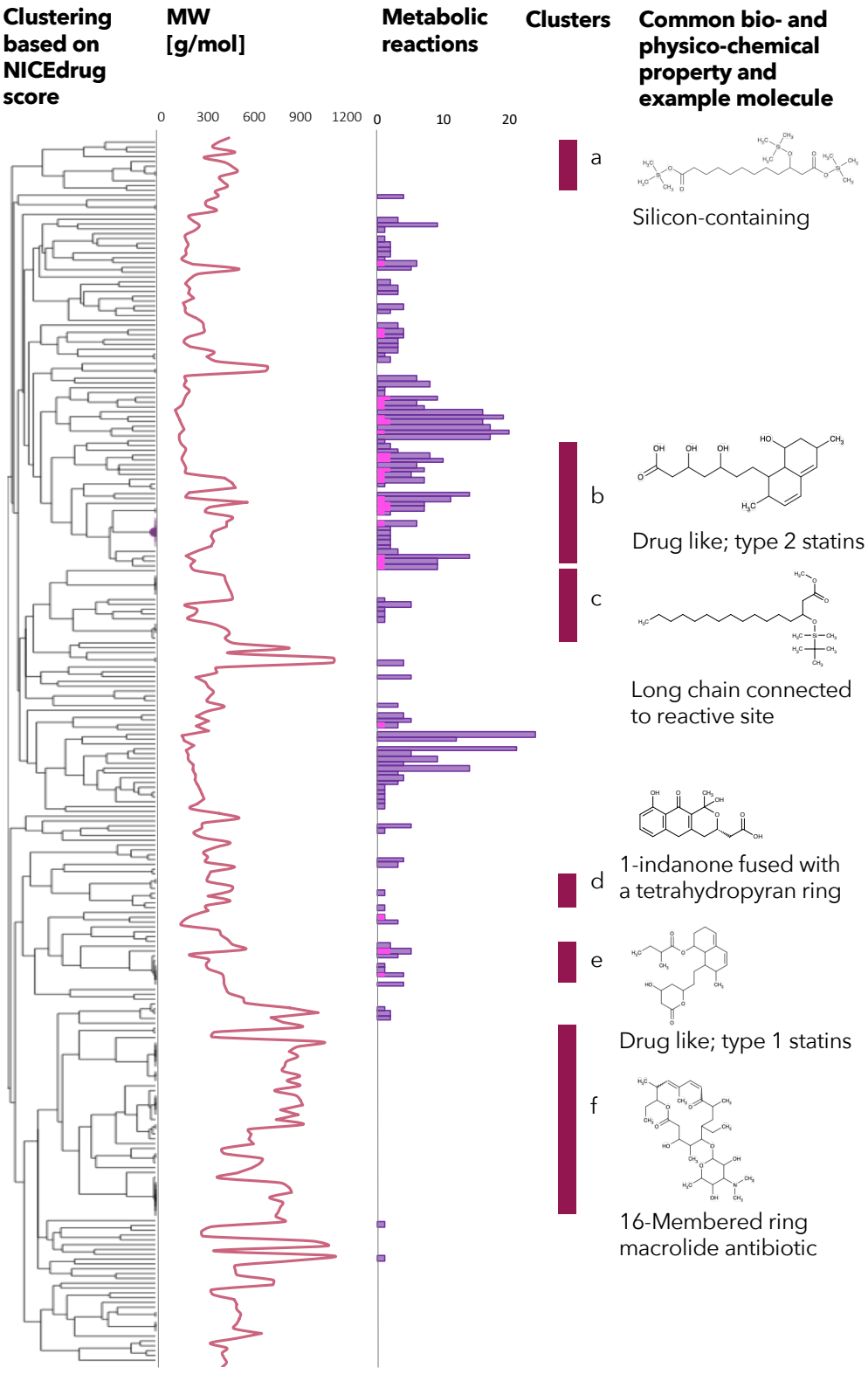


Figure 8.5: Clustering based on NICEdrug score, molecular weight, and reactivity of statin like molecules. Hierarchical clustering based on the NICEdrug score of all molecules in NICEdrug.ch that contain statin reactive site (left). We report the molecules' molecular weight (middle left) and number of drug metabolic reactions or reactions in which these drugs participate (middle). The molecular

weight seems to be inversely correlated with the number of drug metabolic reactions. We highlight six clusters of drugs (a-f, middle right) and an example representative molecule (left). Interestingly, these clusters also group molecules based on bio- or physico-chemical properties: “cluster a” involves a range of silicon-containing chemical molecules, “cluster b” are drug like molecules of type 2 statins, “cluster c” includes chemical molecules with a long chain connected to the reactive site, “cluster d” involves molecules with 1-indanone fused with a tetrahydropyran ring, “cluster e” comprises drug-like molecules of type 1 statins, and “cluster f” are 16-membered ring macrolide antibiotics, related to chapter 6.

Curriculum Vitae

Homa MohammadiPeyhani

Assistant Doctorant

Laboratory of Computational systems biology

École Polytechnique Fédéral de Lausanne (EPFL)

Email: homa.mohammadipeyhani@epfl.ch

Institut des sciences et ingénierie chimiques
EPFL SB ISIC LCSB
CH H4 625 (Bât. CH)
Station 6, CH-1015 Lausanne
Switzerland

Tel.: +41 786969489

Education

2016 – Present | **PhD. in Chemical Engineering/Computational biology, Swiss Federal Institute of Technology (EPFL)-Marie Curie Fellow**

Thesis aim: Design new methods for reaction similarity evaluation, and expanding idea to genome sequence analysis and protein design.

Advisor: Prof. Vassily Hatzimanikatis

Teaching: Introduction to chemical engineering, Bioreactor design and modeling

2014 – 2016 | **M.Sc. in Chemical Engineering/ Control process, Sharif University of Technology**

Thesis:
Ontological study of metabolism to develop an identifications software for metabolic networks.

Adviser: Prof. Ramin Bozorgmehri
Results:
Total GPA of 5.61 on a 6-point scale, Ranked first among all master students in Chemical and

Petroleum engineering department (more than 50 students).

Teaching: fundamentals of Genetics, mass and energy balance, Industrial chemistry II, unit operation, Mass transfer

2010 – 2014 **B.Sc. in Chemical Engineering, Sharif University of Technology**

Thesis: Modeling of Hyaluronic Acid metabolic pathway; using *pichia pastoris* gene regulation and control the fermentation production of different molecular weight Hyaluronic acid.
Adviser: Prof. Shohreh Mashayekhan
Results:
Total GPA of 5.55 on a 6-point scale, Ranked first among all bachelor students in Chemical and Petroleum engineering department (more than 120 students).

Field of Interest

- Bioinformatics,
- Computational biology,
- Protein design and metabolic modeling,
- Drug design,
- Big data analysis,
- Data mining,
- Machine learning,
- Process control,

Awards and Honors

- Best poster award LIMNA Symposium, 2019.
- Best poster award SystemsX.ch conference on system biology, 2018.
- Marie Skłodowska-Curie Fellowship (2016).
- Recipient of the Grant for Graduate Studies from Iran National Elites Foundation (2014-2016).
- Selected as an exceptional talent by Sharif University of Technology and been offered a place at Chemical engineering master program (control and simulation, 2014).
- With Average score18.50 (From20), graduated as the first student of Chemical and Petroleum engineering department among student Class of 2010 (more than 120 students), Sharif University of Technology (2014).
- Ranked 526th among more than 317,600 participants in University Entrance Examination, Iran (2010).

Research Experiences and internships

Summer 2019 **Data analysis in a biological context, SILICO LIFE Company, Braga, Portugal**

- Big data analysis
- Network study
- Database management

2018 **Collaboration on prediction of biosynthesis routes for NRC target compounds, NESTLE Company, Lausanne Switzerland**

- Curation of database of protein sequences
- Predict/design metabolic/catabolic pathways based on known biochemistry

2017-2018

Collaboration on production of adipic acid project, University of Denmark (DTU), Copenhagen, Denmark

- Prediction of novel pathways for adipic acid biosynthesis,
- Proposing enzymes for novel reactions
- Analyzing the feasibility of the predicted pathways by integrating them in the metabolism of a microorganism.

2017

Collaboration on retro biosynthesis project, L'Oréal Company, Paris, France
In-silico metabolic pathway prediction

Entrepreneurship center of Sharif university of technology, Tehran, Iran
Data analysis.

2011-2016

Publications

2020

Jasmin Hafner[1], Homa MohammadiPeyhani[1], Anastasia Sveshnikova, Alan Scheidegger, Vassily Hatzimanikatis, "Updated ATLAS of Biochemistry with new metabolites and improved enzyme prediction power," ACS, Synthetic biology., Accepted.

1 contributed equally

2019

N. Hadadi [1], H. MohammadiPeyhani [1], L. Miskovic, M. Seijo, and V. Hatzimanikatis, "Enzyme annotation for orphan and novel reactions using knowledge of substrate reactive sites," Proc. Natl. Acad. Sci., p. 201818877, Mar. 2019.

1 contributed equally