

Modeling, predicting and mining metabolism at atom-level resolution

Présentée le 18 septembre 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

Jasmin Maria HAFNER

Acceptée sur proposition du jury

Prof. Y. Aye, présidente du jury
Prof. V. Hatzimanikatis, directeur de thèse
Prof. V. De Lorenzo, rapporteur
Prof. J.-L. Faulon, rapporteur
Prof. U. Sauer, rapporteur

Acknowledgements

Writing these acknowledgements after four and a half years of PhD makes me realize how many people contributed to this piece of work, directly and indirectly, by providing guidance and support on many different levels.

First of all, I want to express my deep gratitude towards my supervisor Vassily Hatzi-manikatis. Vassily, you gave me the opportunity to work and to grow, as a scientist and as a human, in your lab. I thank you for the tough challenges as well as for the enlightening teachings, which contributed to shaping the person I am today.

Next, I would like to acknowledge my PhD committee, Prof. Victor de Lorenzo, Prof. Jean-Loup Faulon, and Prof. Uwe Sauer for reading my work, travelling all the way to Lausanne, and sharing with me their criticism and ideas. I am also very thankful to my supervisor at Stanford University, Prof. Christina Smolke, for accepting me in her lab, and for giving me the unique opportunity to collaborate with synthetic biologists on natural product biosynthesis in yeast. At this point, I would also like to acknowledge the private company Firmenich, who provided the financial means for this research exchange.

Furthermore, I want to thank my previous direct supervisor and friend Noushin Hadadi for accepting me as a master student in Vassily's lab six years ago. Noushin, you opened the door for me to new realms of research, to new cultures, and to new scientific challenges. I further want to acknowledge my students for their contributions to this thesis: My first master student, Beatriz Lopes, for her insisting on very complex problems and for not giving up in difficult situations. Bea, thanks for the long discussions and the many mojitos we shared, in Lausanne or in Porto, you became a dear friend to me! I further thank my second master student, Basile Laurent, for his excellent work on biodegradation, as well as for his positive contribution to the atmosphere in the lab with his cheerful character. I also thank my semester students Victor Viterbo, Alan Scheidegger and Claire Stoffel for their valuable contributions on the ATLAS projects. Special thanks go to Adrian Shajkofci, who taught me how to handle the most complicated code structures, how to catch pokémon, and who would always answer my biologist's questions. Adrian, my partner in crime, thank you for making the long nights in lab working on ATLAS or course projects more enjoyable!

I also want to thank my coworkers, in particular Homa Mohammadi-Peyhani, Zhaleh Hosseini, Anastasia Sveshnikova, and Anush Chiappino-Pepe. Girls, team mates, I'm truly thankful that I could always count on you, and that you never (or only rarely) lost patience with

me. Thank you for your insisting when I was wrong, for correcting me, for bringing in new ideas, and for making things happen in the end. Working with you was a wonderfully enriching experience! During my time in California, I also had the pleasure to work with James Payne, who taught me the secrets of yeast engineering and who had to deal with my computational predictions. James, thank you for your infinite patience with me, and for the hours of work you devoted trying to express my predicted enzymes!

Finally, I want to say thank you to the whole LCSB family for the mental support, the good parties and barbecues, the salad club, and the beers at Satellite – Misko, Christine, Maria, Daniel, Pierre, Robin, Asli, Evangelia, Maxime, Sofia, Omid, Liliana as well as Meriç, Tiziano and many others in earlier times. Talking about Satellite: An important *thank you* goes to the association for organizing all the great concerts, and to the amazing people behind the bar – Arthur, Bastien, Jean-Luc, Lionel, Bébé, and everyone else I forgot to mention – big thanks for the countless coffees, beers and smiles!

When things got difficult in the lab, I was lucky to count on the company of people around me at home. During my PhD time, I shared the place called home with an estimated number of twenty roommates, some of which influenced me in particular: Thank you Lauranne and Grég for the wonderful cocktails in the bar downstairs and the smelly French cheese in the fridge, and little Lyra for the love she spread in our home. Also, thank you Stéphanie, Aurélie and Vanessa for the girl power, the coffees and the nights out. A big *thanks* goes to my roomies AJ and Henry in East Palo Alto, CA, for the great welcome and all the craziness in sunny California. In particular, thank you AJ for teaching me rocket science and, more importantly, constantly encouraging me, especially when times were difficult. I also want to say thanks to my current roommates – Coco, Damien, Bastien, Papillon, and previously Angela – for the support, especially in the last, painful phase of writing this thesis. You guys are my family. I don't know how I would have done it without your patience, your care, your cooking (Damdam t'es le meilleur), your psychological support (Bébé) and your sharp analyses of problems of any kind (Coco). Thank you *les bolocs* for the vital energy you bring into my life every day, I love you guys!

I wouldn't have done all of this without the friendship and support of my close friends. To Vivian and Steffi, who started their PhD at same time as me, thank you girls for sharing the PhD experience, for all the nerdy discussions about code, and for the outdoors adventures. Talking about adventures, thank you Grégoire for the many hiking and climbing outings that let me forget about the stressful work. Katinka – one of the very few friends who joins me in jumping into the icy cold Lac Léman in the midst of winter – thank you so much for providing me shelter from the PhD stress in Davos and Alp Flix, and for the many completely improvised trips, hikes, and food experiments. You taught me not to listen to people, but to my heart and my own crazy ideas. Andreas, thank you for being around in the most difficult times, and for having the courage to try to teach me how to drive a car. Camille, thanks for the amazing times working and partying at the Reeds festival, for the nights out in Züri, for the “Dreadies”, and for all the biologist's discussions on plants and their feelings. Sneha, my dear sister, thank you for the all the endless discussions in Nice, Paris, Annecy, or wherever

we happened to meet. You are inspiring me in so many ways. Stéphane, thank you for being there when I needed you, for the coffees and cigarettes, and for being the friend you are to me. Lucie, thank you for the love and friendship, for the many drinks and discussions that were needed to solve the small and the bigger problems. I could always count on you.

Finally, I want to thank my uncle Herbie and his wife Bea, as well as my grandmother Ursula and my grandaunt Marianne, for the support during my studies and for the inspiring discussion across generations. Last but not least, I want to say thank you to my wonderful, loving family: My parents Brigitta and Reto, and my brother Samuel (Brüderhärzli). I don't have words to describe what you gave to me, but I want to let you know that I'm infinitely grateful for the love, care, and everything.

Lausanne, March 2020

"The day that you stop running is the day that you arrive"

– Morcheeba in "Enjoy The Ride"

Abstract

Living organisms can catalyze many thousands of biochemical reactions that they use to convert energy and matter, which provides them with the essentials for life. The sum of these chemical reactions happening in an organism is called metabolism. Understanding metabolism is crucial to elucidate the fundamental principles of biology, and further to enable us to redesign it for the sustainable, biosynthetic production of bio-based fuels, commodity chemicals and medicines. Mathematical models are essential to organize and understand the complexity of metabolism. They usually represent metabolism as a network of reactions, but they tend to neglect the exact molecular structure of the metabolites. To redesign metabolic reactions, however, a mechanistic understanding of metabolic reactions and their catalysts, proteins called enzymes, is essential.

In this work, a mathematical description of enzymatic reaction mechanism, called generalized reaction rules, is applied to computationally simulate and predict metabolic processes at the level of atoms. Each reaction rule describes the catalytic activity of an enzyme, or a group of enzymes, at the mechanistic level by encoding the rearrangement of atoms in the reaction. The reaction rules are called "generalized", because they mimic the ability of a single enzyme to catalyze multiple reactions by acting on a range of substrates.

Using these reaction rules, we first developed a computational representation of metabolism that allowed tracking single atoms throughout complex metabolic reaction networks. The principle of atom-tracking was then used to develop a graph-theory based method to represent and analyze metabolic networks, and to reliably identify metabolic pathways for the biosynthesis of chemicals. Next, we applied the generalized reaction rules to predict all possible novel, hypothetical reactions from known biological compounds, and we stored the five million generated novel reactions in a database called ATLAS. Finally, the developed tools and resources were applied to specific engineering and research problems, such as the biosynthetic pathway design for the biofuel bisabolene and the plastic precursor 1,4-butanediol. We further predicted a biosynthesis route for the pharmaceutical tetrahydropalmatine and engineered a yeast strain to produce it. Finally, we show that our tools can be used to mine available genome sequences to find organisms that can degrade xenobiotics.

Our findings suggest that the atom-level representation of metabolism can greatly contribute to its understanding, exploration and prediction. Given the complexity of atom-level modeling of metabolic processes, we propose metrics that can approximate the atom-level information to conserve the information at the level of big, hypothetical metabolic

networks like ATLAS. This database plus the developed pathway search techniques form a valuable resource for scientists to help characterizing unknown biosynthesis pathways towards secondary metabolites, and for metabolic engineers to design novel bioproduction pathways for chemicals. Hopefully, these considerations will contribute to a better understanding of metabolism, advance the exploration of the bioproduction of drugs and other valuable molecules, and accelerate metabolic engineering efforts to realize the switch from a petroleum-based chemical industry towards a more sustainable, bio-based production of society's chemical needs.

Keywords

Computational biology, metabolic modeling, biochemical networks, atom-mapping, reaction prediction, enzyme promiscuity, pathway search, metabolic engineering

Résumé

Les organismes vivants ont la capacité de catalyser des milliers de réactions chimiques qu'ils utilisent lors de la transformation de l'énergie et de la matière, leur fournissant les éléments essentiels à la vie. L'ensemble de toutes ces réactions biochimiques est appelé métabolisme. La connaissance du métabolisme est non seulement déterminante pour la bonne compréhension des principes fondamentaux de la biologie, mais aussi le pré-requis dans la modification d'organismes afin de produire des biocarburants, des produits chimiques et des médicaments de façon biosynthétique. Des modèles mathématiques sont essentiels pour organiser et comprendre la complexité du métabolisme. Dans ces modèles, les processus métaboliques sont représentés par des réseaux de réactions biochimiques, mais ces représentations ont tendance à négliger les structures moléculaires exactes des métabolites. Pour construire et modifier des réactions métaboliques à des fins d'ingénierie biologique, il est essentiel de comprendre ces réactions et leurs catalyseurs, les enzymes, au niveau mécanistique.

Dans ce travail, nous avons utilisé des descriptions mathématiques des mécanismes de réaction enzymatiques, appelés règles de réaction généralisées, pour les appliquer à la simulation et prédiction de processus métaboliques au niveau atomique. Chaque règle de réaction définit l'action d'une enzyme au niveau mécanistique par la description du réarrangement des atomes accompli par l'enzyme. De plus, ces règles sont appelées "généralisées" parce qu'elles reproduisent la capacité d'une seule enzyme à catalyser plusieurs réactions différentes en agissant sur une série de substrats.

Sur la base de ces règles de réactions, nous avons créé un modèle computationnel du métabolisme, permettant le suivi individuel des atomes à travers la complexité des réactions métaboliques. Le traçage des atomes a ensuite été utilisé pour développer une méthode basée sur la théorie des graphes pour analyser de manière systématique des réseaux métaboliques, et y identifier de façon fiable des voies métaboliques pour la biosynthèse des molécules chimiques. Nous avons ensuite utilisé les règles de réaction pour la prédiction de nouvelles réactions hypothétiques. En appliquant celles-ci à des composés biologiques, nous avons créé une base de données de cinq millions de nouvelles réactions, appelée ATLAS. Cette ressource peut aider les scientifiques à caractériser les voies biosynthétiques des molécules biologiques, ainsi qu'être utilisée dans le cadre de la bio-ingénierie pour identifier des voies de synthèse biologiques de composés chimiques. Finalement, nos outils ont été appliqués à des problèmes spécifiques d'ingénierie et de recherche, comme le design des voies biosynthétiques d'un biocarburant et d'un précurseur de la synthèse de plastique. De plus, ces outils nous ont permis de créer une levure capable de produire le composé

pharmaceutique tétrahydropalmatine, ainsi que d'identifier des organismes aptes à la biodégradation de xénobiotiques.

Ce travail suggère que la représentation du métabolisme au niveau atomique contribue à son analyse, son exploration et à sa prédiction. Nous espérons que ces considérations vont participer à une meilleure compréhension du métabolisme, avancer la biosynthèse de composés de valeur industrielle ou pharmaceutique, et accélérer les efforts d'un changement d'une industrie basée sur la pétrochimie, vers une production biologique et durable des composés chimiques.

Mots-clés

Biologie computationnelle, modélisation du métabolisme, réseaux biochimiques, cartographie des atomes, prédiction de réactions, promiscuité des enzymes, recherche de voies métaboliques, ingénierie métabolique

Contents

Acknowledgements	iii
Abstract.....	vii
Keywords.....	viii
Résumé.....	ix
Mots-clés.....	x
Contents	xi
List of Figures.....	xvii
List of Tables.....	xxiii
List of Abbreviations	xxviii
Chapter 1 Introduction	31
1.1 Nature’s language	32
1.1.1 Systems biology	33
1.1.2 Data in biology: the “omics” era	33
1.1.3 Synthetic biology and metabolic engineering.....	34
1.2 Metabolism	35
1.2.1 Metabolic modeling.....	36
1.2.2 Knowledge gaps in metabolism	37
1.3 Metabolism at the level of atoms	39
1.3.1 Enzymes <i>in silico</i>	40
1.3.2 Atom-level resolution of reaction mechanism	42
1.3.3 Promiscuity-based prediction of biochemical reactions.....	42
1.3.4 Predicting putative enzymes for novel reactions	43
1.4 This thesis.....	43
References.....	45
Chapter 2 Atom-level resolution of metabolic networks	49

2.1	Quantification of cellular metabolic fluxes	49
2.1.1	^{13}C -Metabolic flux analysis (^{13}C -MFA)	50
2.1.2	Optimization-based methods to estimate flux ranges.....	51
2.1.3	FBA <i>versus</i> MFA.....	52
2.2	Tracing single atoms through reactions, pathways and networks	53
2.2.1	Atom-mapped biochemical reactions	54
2.2.2	Atom-mapped metabolic pathways.....	55
2.2.3	Tracing atoms in the substrate through metabolic networks	55
2.3	Atom-level modeling of <i>E. coli</i>	57
2.3.1	Reduction of the <i>E. coli</i> GEM.....	57
2.3.2	Thermodynamic analysis of redEcoli.....	58
2.3.3	Curation of redEcoli with reaction mechanisms	59
2.3.4	Atom-mapped core metabolism of redEcoli	60
2.3.5	Atom-mapped biomass production pathways.....	61
2.3.6	The redEcoli atom-level network.....	65
2.3.7	Construction of an atom-level, stoichiometric hybrid model of redEcoli	66
2.3.8	Refining flux ranges by incorporating experimental ^{13}C distributions ..	67
2.3.9	Conclusions and Outlook	68
2.4	Tracking substrate utilization in the malaria parasite	69
2.4.1	Tracing substrate atoms in a reduced model of <i>P. falciparum</i>	70
2.4.2	Conclusions	71
2.5	Conclusion and outlook.....	72
	References	73
	Chapter 3 Atom-conserving metabolic pathway search.....	77
3.1	The quest for metabolic pathways.....	77
3.1.1	Graph representation of metabolism	78
3.1.2	Atom-conserving pathway search for large biochemical networks	79
3.2	The NICEpath method	80
3.2.1	Biochemically correct atom-mapping with BNICE.ch.....	80
3.2.2	Calculation of weighted reactant-product pairs	81

3.2.3	Assigning mechanisms to biochemical reactions from the KEGG reference network	82
3.2.4	Graph representation of biochemical networks	82
3.2.5	Finding metabolic pathways with graph search	83
3.2.6	Network analysis	83
3.2.7	Software	83
3.3	Results and discussion	84
3.3.1	The CAR captures the main biotransformations	84
3.3.2	Graph-theoretical analysis of metabolic networks	85
3.3.3	Finding biologically relevant pathways with NICEpath	86
3.3.4	Limitations and future challenges	89
3.4	Conclusion	90
	References	91
	Chapter 4 ATLASx - Databases for predictive biochemistry	93
4.1	“Dark matter” in metabolism	93
4.1.1	Cheminformatic approaches	94
4.1.2	The ATLAS of Biochemistry	95
4.2	The ATLAS workflow	97
4.2.1	Database curation	98
4.2.2	Reconstruction of known reactions	99
4.2.3	Prediction of novel reactions	99
4.2.4	Reaction annotation and analysis	99
4.2.5	ATLAS web interface	100
4.3	The update - ATLAS 2018	101
4.3.1	Reconstruction of known reactions	101
4.3.2	Validation of novel ATLAS reactions	102
4.3.3	ATLAS 2018 statistics	102
4.4	bioATLAS and chemATLAS - reactions emerging from biological and bioactive compounds	105
4.4.1	New tools & methods	106
4.4.2	Collecting biochemical data	107
4.4.3	Analysis of reactive sites	109

4.4.4	Prediction of novel reactions	110
4.4.5	Network analysis	110
4.5	Conclusion and outlook.....	113
References	115
Chapter 5	Applications: Predicting biotransformations with cheminformatic tools ..	119
5.1	Retrobiosynthesis.....	119
5.1.1	Iterative network generation	122
5.1.2	Pathway search within hypothetical metabolic network.....	123
5.1.3	Stoichiometric and thermodynamic pathway feasibility	124
5.1.4	Finding enzymes for predicted reactions.....	124
5.1.5	Ranking, visualization and availability.....	125
5.2	Retrobiosynthesis for 1,4-butanediol and bisabolene	127
5.2.1	Generation of biochemical reaction network around target compounds	128
5.2.2	Finding pathways to host precursors.....	129
5.2.3	Stoichiometric, thermodynamic and biocatalytic pathway evaluation	131
5.2.4	Pathway ranking, visualization and comparison.....	133
5.2.5	Conclusion.....	135
5.3	Exploring the chemodiversity around the noscapine pathway	136
5.3.1	Computational expansion of the noscapine pathway.....	138
5.3.2	Ranking candidate compounds by popularity.....	140
5.3.3	Construction of biosynthetic pathways to target compounds.....	141
5.3.4	Selection of candidate enzymes for tetrahydropalmatine bioproduction	145
5.3.5	<i>In vitro</i> and <i>in vivo</i> bioproduction of tetrahydropalmatine	147
5.3.6	Conclusion.....	150
5.4	Predicting potential biodegradation of xenobiotics.....	151
5.4.1	Predictive biodegradation workflow.....	152
5.4.2	Evaluation of biodegradation for six xenobiotics.....	153
5.4.3	Conclusions	155
References	156

Chapter 6	Conclusions and perspectives	165
6.1	Conclusions	165
6.2	Future perspectives	166
6.2.1	Towards non-model organisms	166
6.2.2	Next-generation pathway design.....	167
References.....		168
Appendix		169
	Table A1: Lumped reactions for 20 BBBs and their reconstruction in iAM.NICE	169
	Table A2: Reactant pairs in KEGG with non-zero CAR and corresponding RPAIR annotation	169
	Table A3: Noscapine pathway derivatives ranked by popularity	169
Curriculum Vitae		170

List of Figures

Figure 1.1: Different levels of unknowns in metabolism.	38
Figure 1.2: Computational representation of the biochemical actors in BNICE.ch.	41
Figure 1.3: Overview on the different levels of metabolism addressed in this thesis. Red numbers indicate respective Chapters and Subchapters.	44
Figure 2.1: The ^{13}C -MFA workflow.	50
Figure 2.2: Generalized reaction rules represent the action of a substrate-promiscuous enzyme at atom-level resolution.	54
Figure 2.3: All the carbon atoms of glucose are traced simultaneously through the first four reaction steps of glycolysis in four generations.	55
Figure 2.4: Tracing the carbon atom at position four of glucose through parts of the central carbon metabolism illustrates the generation of carbon-labeled networks in iAM.NICE (adapted from Hadadi <i>et al.</i> ¹⁶).	56
Figure 2.5: The ^{13}C -FBA workflow explains the creation of a reduced hybrid model in five steps, plus a final step for analysis and model validation. Abbreviations: genome-scale model (GEM), Thermodynamic Flux Analysis (TFA), flux directionality profile (FDP), bidirectional reactions (BDRs), amino acids (AA), biomass building blocks (BBBs). The third-level EC numbers in step three represent enzymatic reaction rules, and LUMP_1 stands for a lumped reaction rule converting core precursors into BBBs.	58
Figure 2.6: Carbon origins and labeled subnetworks for methionine (A) and tryptophan (B). Different atom positions in the amino acids are indicated with different colors. The number of reaction steps between the BBB and its labeled precursor is given. For the subnetwork visualization, different colors are used to mark the reaction that carry a label in the different subnetworks (colored edges) as well as the corresponding precursors (colored frames). (A) Blue arrows represent the subnetworks S1_met, S4_met, S5_met and S7_met, red by S2_met, and green and turquoise by S3_met, S6_met and S8_met. The turquoise	

subnetwork is an alternative reaction path that does is redundant with the green route in terms of labeling and final lumped stoichiometry. (B) The blue reaction maps to subnetwork S1_trp, and the pink reactions show the alternative used in S2_trp. Green and yellow arrows show alternative routes without influence on the labeling pattern. The short names of the metabolites and enzymes match the identifiers from iJO1366³⁶.65

Figure 2.7: Schematic of the hybrid model of redEcoli. Thick grey arrows represent labeled reactions, thin black arrows represent unlabeled reactions, and dashed arrows represent lumped reactions. AA: amino acid, BBB: biomass building block, glc: glucose, g6p: glucose 6-phosphate, f6p: fructose 6-phosphate, PGI: phosphoglucoisomerase, glc_ex: glucose exchange flux, ala: alanine, prec: precursor..... 67

Figure 2.8: Substrate metabolites can substitute each other in the media in iPfa. Each essential substructure, or moiety, can be obtained from a range of alternative substrates. Adapted from Chiappino-Pepe et al.⁴⁴.70

Figure 2.9: Example output of a substrate utilization study on guanosine monophosphate (GMP). Different colors designate the corresponding atomic positions in the molecular structures of the BBBs. THF: Tetrahydrofolate, GTP: Guanosine triphosphate.72

Figure 3.1: The workflow of the pathway search is divided in two parts. The first two steps (left) describe the atom weighting of the network from atom-mapped reactions. In this study, steps 1 and 2 are performed by BNICE.ch. Steps 3 and 4 (right), implemented in NICEpath, take the atom-weighted network as an input to create a searchable graph structure and finally apply a Yen's k-shortest pathway search.80

Figure 3.2: Example of relation between KEGG RPAIRs and the CAR value in a biochemical reaction. (A) Alcohol dehydrogenase, (B) decarboxylation reaction.84

Figure 3.3: (Left) The ROC curve shows the prediction of KEGG RPAIRs of type "main" by CAR score from BNICE.ch. (Right) The trade-off between specificity (blue) and sensitivity (red). The Youden's index (yellow) reaches its maximum (0.66) at a CAR value of 0.34.85

Figure 3.4: The pathways from Table 3.1 with index numbers 1, 2 and 5 connecting tyrosine and caffeate are visualized in detail for comparison. For each biotransformation, the CAR value as well as the default distance (d) are indicated.87

Figure 3.5: Comparison of two pathways (A and D*) from the pathway search connecting tyrosine to syringin. For each biotransformation, the

CAR value as well as the default distances for each transformation are indicated. d(dflt): default distance, d(sqrt): square root transformation, d(exp): exponential transformation. 89

Figure 4.1: The biochemical reaction network of the ATLAS of biochemistry. Reactions are color-coded according to their pathway annotations in KEGG. The network has been drawn in the open-source graph tool Gephi²⁴. 96

Figure 4.2: Defining the scope of different compound spaces and their associated ATLAS projects. 96

Figure 4.3: Overview on the general ATLAS workflow 98

Figure 4.4: 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) rat109456 along with its most similar reaction and candidate enzyme, predicted by an earlier version of BridgIT to calculate structural reaction similarities. (right) rat109456 in ATLAS 2018 is now cataloged in KEGG as R11332 with EC 5.3.1.33. Two alternative enzyme candidates are proposed by the updated version of BridgIT. 104

Figure 4.5: Different types of reactions in ATLAS derived from of the biological and bioactive compound space. 106

Figure 4.6: (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 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[tricarboxyl(η 5-cyclopentadienyl)molybdenum](Mo—Mo). 109

Figure 4.7: (A) Visual overview on different statistics and network properties calculated for bioDB, bioATLAS and chemATLAS. (B) Size distribution of disconnected components in the network of each of the three database scopes. 112

Figure 5.1: Schematic of BNICE.ch-based retrobiosynthesis workflow. A hypothetical biochemical network is expanded around the target compound (red dot). Pathways connecting the host metabolism to the target (red path) are retrieved and evaluated for stoichiometric, thermodynamic and enzymatic feasibility. 120

Figure 5.2: The BNICE.ch network generation process. 123

Figure 5.3: Compounds produced in each BNICE.ch iteration during the network generation process for 1,4-BDO (blue) and bisabolene (orange).

Solid lines indicate the total number of compounds, and the dashed lines show the number of biological KEGG compounds.	129
Figure 5.4: Distribution of pathway lengths for 1,4-BDO in <i>E. coli</i> and for bisabolene in <i>E. coli</i> and <i>S. cerevisiae</i>	130
Figure 5.5: Distribution of pathway length after stoichiometric and thermodynamic curation, respectively. Precursor compounds of the shortest pathways are listed for each analysis.	132
Figure 5.6: Comparison of the three synthetic 1,4-BDO bioproduction pathways in <i>E. coli</i> with two of the top-ranked pathways in this study. Protonation states and corresponding names of compounds originating from BNICE.ch are not corrected for pH 7. For example, 4-hydroxybutanoic acid corresponds to 4-hydroxybutyrate in standard biological conditions.	134
Figure 5.7: Biosynthesis pathways for bisabolene in nature, and top-ranked pathways in <i>E. coli</i> and yeast. Abbreviations: Inorganic phosphate (Pi) and diphosphate (PPi).	135
Figure 5.8: Overall workflow. Left: Applied design-build-test cycle. Right: 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.	137
Figure 5.9: Visualization of the expanded biosynthesis network of the noscapine pathway. The nodes and edges drawn in red shows 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 has been 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. The free graph visualization tool Gephi was used for network visualization ⁷⁸	140
Figure 5.10: Metabolic pathway to (S)-tetrahydropalmatine from (S)-norcoclaurine in yeast. For each reaction, the enzyme identifier and the EC number are indicated. In addition, the similarity of each reaction with respect to the proposed biosynthetic step for tetrahydropalmatine is	

indicated with the BridgIT score (bold red) obtained by BridgIT analysis.	146
Figure 5.11: Tetrahydropalmitine production in yeast transformed with different OMT-encoding plasmids after 5 days growth. +pCS2812: low- copy number plasmid, +pC2952: high-copy number plasmid. MRM: Multiple Reaction Monitoring counts.	148
Figure 5.12: (A) Screening methyltransferases in tetrahydropalmitine pathway for enzymatic activity on (S)-tetrahydrocolumbamine. (B) Quantification of enzyme activity, comparing CjOMT, Ps6OMT, yPsS9OMT, CNMT. The product of CNMT is not tetrahydropalmitine, but presumably the N-methylated product of tetrahydrocolumbamine (structure shown). EIC: Extracted Ion Counts.	150
Figure 5.13: Workflow to assess the biodegradability of xenobiotics and to identify organisms with the potential capacity to degrade the xenobiotic of interest.	153
Figure 5.14: (A) Compounds chosen as input for the proposed biodegradation workflow. (B) Network generation around six xenobiotics using BNICE.ch. The number of compounds is indicated after each iteration in the network generation process. Compounds colored in blue map to the left scale, and compounds colored in green map to the right scale.	154

List of Tables

Figure 1.1: Different levels of unknowns in metabolism.	38
Figure 1.2: Computational representation of the biochemical actors in BNICE.ch.	41
Figure 1.3: Overview on the different levels of metabolism addressed in this thesis. Red numbers indicate respective Chapters and Subchapters.	44
Figure 2.1: The ¹³ C-MFA workflow.	50
Figure 2.2: Generalized reaction rules represent the action of a substrate-promiscuous enzyme at atom-level resolution.	54
Figure 2.3: All the carbon atoms of glucose are traced simultaneously through the first four reaction steps of glycolysis in four generations.	55
Figure 2.4: Tracing the carbon atom at position four of glucose through parts of the central carbon metabolism illustrates the generation of carbon-labeled networks in iAM.NICE (adapted from Hadadi <i>et al.</i> ¹⁶).	56
Figure 2.5: The ¹³ C-FBA workflow explains the creation of a reduced hybrid model in five steps, plus a final step for analysis and model validation. Abbreviations: genome-scale model (GEM), Thermodynamic Flux Analysis (TFA), flux directionality profile (FDP), bidirectional reactions (BDRs), amino acids (AA), biomass building blocks (BBBs). The third-level EC numbers in step three represent enzymatic reaction rules, and LUMP_1 stands for a lumped reaction rule converting core precursors into BBBs.	58
Figure 2.6: Carbon origins and labeled subnetworks for methionine (A) and tryptophan (B). Different atom positions in the amino acids are indicated with different colors. The number of reaction steps between the BBB and its labeled precursor is given. For the subnetwork visualization, different colors are used to mark the reaction that carry a label in the different subnetworks (colored edges) as well as the corresponding precursors (colored frames). (A) Blue arrows represent the subnetworks S1_met, S4_met, S5_met and S7_met, red by S2_met, and green and turquoise by S3_met, S6_met and S8_met. The turquoise subnetwork is an alternative reaction path that does is redundant with	

the green route in terms of labeling and final lumped stoichiometry. (B) The blue reaction maps to subnetwork S1_trp, and the pink reactions show the alternative used in S2_trp. Green and yellow arrows show alternative routes without influence on the labeling pattern. The short names of the metabolites and enzymes match the identifiers from iJO1366³⁶. 65

Figure 2.7: Schematic of the hybrid model of redEcoli. Thick grey arrows represent labeled reactions, thin black arrows represent unlabeled reactions, and dashed arrows represent lumped reactions. AA: amino acid, BBB: biomass building block, glc: glucose, g6p: glucose 6-phosphate, f6p: fructose 6-phosphate, PGI: phosphoglucoisomerase, glc_ex: glucose exchange flux, ala: alanine, prec: precursor..... 67

Figure 2.8: Substrate metabolites can substitute each other in the media in iPfa. Each essential substructure, or moiety, can be obtained from a range of alternative substrates. Adapted from Chiappino-Pepe et al.⁴⁴. 70

Figure 2.9: Example output of a substrate utilization study on guanosine monophosphate (GMP). Different colors designate the corresponding atomic positions in the molecular structures of the BBBs. THF: Tetrahydrofolate, GTP: Guanosine triphosphate. 72

Figure 3.1: The workflow of the pathway search is divided in two parts. The first two steps (left) describe the atom weighting of the network from atom-mapped reactions. In this study, steps 1 and 2 are performed by BNICE.ch. Steps 3 and 4 (right), implemented in NICEpath, take the atom-weighted network as an input to create a searchable graph structure and finally apply a Yen's k-shortest pathway search. 80

Figure 3.2: Example of relation between KEGG RPAIRs and the CAR value in a biochemical reaction. (A) Alcohol dehydrogenase, (B) decarboxylation reaction. 84

Figure 3.3: (Left) The ROC curve shows the prediction of KEGG RPAIRs of type "main" by CAR score from BNICE.ch. (Right) The trade-off between specificity (blue) and sensitivity (red). The Youden's index (yellow) reaches its maximum (0.66) at a CAR value of 0.34. 85

Figure 3.4: The pathways from Table 3.1 with index numbers 1, 2 and 5 connecting tyrosine and caffeate are visualized in detail for comparison. For each biotransformation, the CAR value as well as the default distance (d) are indicated. 87

Figure 3.5: Comparison of two pathways (A and D*) from the pathway search connecting tyrosine to syringin. For each biotransformation, the CAR value as well as the default distances for each transformation are

indicated. d(dflt): default distance, d(sqrt): square root transformation, d(exp): exponential transformation.	89
Figure 4.1: The biochemical reaction network of the ATLAS of biochemistry. Reactions are color-coded according to their pathway annotations in KEGG. The network has been drawn in the open-source graph tool Gephi ²⁴	96
Figure 4.2: Defining the scope of different compound spaces and their associated ATLAS projects.....	96
Figure 4.3: Overview on the general ATLAS workflow	98
Figure 4.4: 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) rat109456 along with its most similar reaction and candidate enzyme, predicted by an earlier version of BridgIT to calculate structural reaction similarities. (right) rat109456 in ATLAS 2018 is now cataloged in KEGG as R11332 with EC 5.3.1.33. Two alternative enzyme candidates are proposed by the updated version of BridgIT.	104
Figure 4.5: Different types of reactions in ATLAS derived from of the biological and bioactive compound space.	106
Figure 4.6: (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 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[tricarboxyl(η 5-cyclopentadienyl)molybdenum](Mo—Mo).	109
Figure 4.7: (A) Visual overview on different statistics and network properties calculated for bioDB, bioATLAS and chemATLAS. (B) Size distribution of disconnected components in the network of each of the three database scopes.	112
Figure 5.1: Schematic of BNICE.ch-based retrobiosynthesis workflow. A hypothetical biochemical network is expanded around the target compound (red dot). Pathways connecting the host metabolism to the target (red path) are retrieved and evaluated for stoichiometric, thermodynamic and enzymatic feasibility.	120
Figure 5.2: The BNICE.ch network generation process.	123
Figure 5.3: Compounds produced in each BNICE.ch iteration during the network generation process for 1,4-BDO (blue) and bisabolene (orange). Solid lines indicate the total number of compounds, and the dashed lines show the number of biological KEGG compounds.....	129

Figure 5.4: Distribution of pathway lengths for 1,4-BDO in <i>E. coli</i> and for bisabolene in <i>E. coli</i> and <i>S. cerevisiae</i>	130
Figure 5.5: Distribution of pathway length after stoichiometric and thermodynamic curation, respectively. Precursor compounds of the shortest pathways are listed for each analysis.....	132
Figure 5.6: Comparison of the three synthetic 1,4-BDO bioproduction pathways in <i>E. coli</i> with two of the top-ranked pathways in this study. Protonation states and corresponding names of compounds originating from BNICE.ch are not corrected for pH 7. For example, 4-hydroxybutanoic acid corresponds to 4-hydroxybutyrate in standard biological conditions.....	134
Figure 5.7: Biosynthesis pathways for bisabolene in nature, and top-ranked pathways in <i>E. coli</i> and yeast. Abbreviations: Inorganic phosphate (Pi) and diphosphate (PPi).	135
Figure 5.8: Overall workflow. Left: Applied design-build-test cycle. Right: 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.....	137
Figure 5.9: Visualization of the expanded biosynthesis network of the noscapine pathway. The nodes and edges drawn in red shows 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 has been 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. The free graph visualization tool Gephi was used for network visualization ⁷⁸	140
Figure 5.10: Metabolic pathway to (S)-tetrahydropalmatine from (S)-norcoclaurine in yeast. For each reaction, the enzyme identifier and the EC number are indicated. In addition, the similarity of each reaction with respect to the proposed biosynthetic step for tetrahydropalmatine is indicated with the BridgIT score (bold red) obtained by BridgIT analysis.	146

Figure 5.11: Tetrahydropalmitine production in yeast transformed with different OMT-encoding plasmids after 5 days growth. +pCS2812: low-copy number plasmid, +pC2952: high-copy number plasmid. MRM: Multiple Reaction Monitoring counts.	148
Figure 5.12: (A) Screening methyltransferases in tetrahydropalmitine pathway for enzymatic activity on (S)-tetrahydrocolumbamine. (B) Quantification of enzyme activity, comparing CjOMT, Ps6OMT, yPsS9OMT, CNMT. The product of CNMT is not tetrahydropalmitine, but presumably the N-methylated product of tetrahydrocolumbamine (structure shown). EIC: Extracted Ion Counts.	150
Figure 5.13: Workflow to assess the biodegradability of xenobiotics and to identify organisms with the potential capacity to degrade the xenobiotic of interest.	153
Figure 5.14: (A) Compounds chosen as input for the proposed biodegradation workflow. (B) Network generation around six xenobiotics using BNICE.ch. The number of compounds is indicated after each iteration in the network generation process. Compounds colored in blue map to the left scale, and compounds colored in green map to the right scale.	154

List of Abbreviations

1,4-BDO	1,4-butanediol
ACALD	Acetaldehyde dehydrogenase
ATP	Adenosine triphosphate
BBB	Biomass building block
BDR	Bi-directional reaction
BIA	Benzylisoquinoline alkaloid
BNICE(.ch)	Biochemical Network Integrated Computational Explorer
CAR	Conserved Atom Ratio
CO₂	Carbon dioxide
CoA	Coenzyme A
DNA	Deoxyribonucleic acid
EC	Enzyme Commission
FBA	Flux Balance Analysis
FDP	Flux directionality profile
FPP	Farnesyl diphosphate
FUM	Fumarase
FVA	Flux Variability Analysis
GC-MS	Gas chromatography-mass spectrometry
GCM	Group Contribution Method
GEM	Genome-scale model
GLNS	Glutamine synthetase
GPR	Gene-Protein-Reaction (association)
KEGG	Kyoto Encyclopedia of Genes and Genomes
MDH	malate dehydrogenase
METS	Methionine synthase
MFA	Metabolic Flux Analysis
NAD(P)	Nicotinamide adenine dinucleotide (phosphate)
OMT	O-methyl transferase
ORF	Open reading frame
PET	Polyethylene terephthalate
PGI	Phosphoglucose isomerase
Pi	Inorganic phosphate
PNP	Plant natural product
PPI	Inorganic diphosphate
RNA	Ribonucleic acid
RPE	Ribulose 5-phosphate 3-epimerase
TALA	Transaldolase

TCA cycle	Tricarboxylic acid cycle
TFA	Thermodynamics-based Flux Analysis
THCB	Tetrahydrocolumbamine
TKT	Transketolase
TPI	Triose-phosphate isomerase
TRPAS2	Tryptophanase
TRPS2	Tryptophan synthase

Chapter 1 Introduction

Our societies today heavily rely on petroleum as a source of energy and everyday products. Burning fossil fuels makes it possible for us to cheaply drive cars, heat our homes, cook food, to name just a few. On top of that, a plethora of products are chemically synthesized from petroleum-based feedstocks: Many pharmaceuticals, industrially relevant chemicals (*e.g.*, paint, glue, solvents), beauty products (*e.g.*, creams, make-up) are chemically synthesized from petroleum-based precursors, also called commodity chemicals. To sum it up, we heavily rely on petroleum – but unfortunately, there are many problems associated with our dependence on the black gold.

While petroleum is a cheap and convenient source of energy, burning it carbon dioxide (CO₂) to our atmosphere, where it accumulates and causes the rise of global temperatures; It is therefore the major driver of anthropogenic climate change. In the case the raw petroleum is not used to produce fuels, it can be chemically refined to bulk chemicals used in petrochemistry, which are used as precursors in petrochemical industry. Chemical synthesis, on their side, make use of toxic catalysts that can pollute the environment if not properly contained. Finally, other than its negative effect on climate and environment and its limited supply, relying on petroleum is problematic from a geopolitical point of view: Petroleum occurs only in specific regions of the globe, and due to the heavy dependence of our societies on the material, petroleum-rich regions are often politically contested and many of them have been the showplace of territorial conflicts in the recent past. Decentralizing our sourcing of energy and bulk chemicals is therefore crucial to move towards more sustainable societies in an intact environment¹.

But is there an alternative to petroleum? Petroleum has been produced over thousands of years from dead organisms such as algae and zooplankton, which have been buried in sedimentary rock, decomposed by bacteria under anaerobic conditions and fossilized under high heat and pressure. Its original source is therefore biomass. If we could source fuels, bulk chemicals and their derivatives directly from biomass, we could decrease the release of additional CO₂ into the atmosphere. One promising solution to achieve the shift from petroleum to biomass is using microbes as miniature chemical factories to convert organic material such as plant waste products into high-value chemicals such as fuels, bulk chemicals and specialty chemicals. Microbes such as yeast or lactic acid bacteria have been used in by humanity for thousands of years to refine food and to save it from perishing through fermentation. For example, the fungus yeast converts sugars to ethanol in the production of wine and beer, and to bubble of CO₂ in bread. Using microbes to refine a biomass source is called biotechnology, a process that can be used to provide green fuels and to sustainably

produce bulk and fine chemicals. To convince microbes to do this job for us however, we need to exactly tell them how to do it. And if we want to *talk* to them, we first need to learn their *language*.

1.1 Nature's language

Deciphering the underlying principles of the biological world is an ongoing quest, and we are still far away from having an understanding comparable to other disciplines such as physics, whose principles can be used to build stable bridges, create computers, or design rockets that go to mars. Biology turned out to be a lot more complicated: Living organisms eat, excrete, grow, reproduce, and die. They can take a mindboggling number of shapes and show behaviors beyond our imagination – and most strangely, we are part of them. First attempts to organize living things resulted in the classification of organisms by Carl Linnaeus in 1758². One century later, Charles Darwin hypothesized that the classes proposed by Linnaeus have been formed through evolution from a common ancestor³, which, at the very beginning, was probably some simple, unicellular being that already did what all biological entities do today – eat, excrete, grow, reproduce, die. However, the molecular mechanism of evolution was not known until Watson and Crick unveiled the genetic code, written on long chains of molecules called DNA. DNA turned out to harbor all the information necessary to build an organism, to keep it alive, and to do all the things that organisms do – climb on trees, be trees, degrade compost, photosynthesize, make you sick, or fly around glowing after sunset.

The DNA encodes how to do all of this: The entities of the genetic code are called *genes*, and they are transcribed into messenger molecules called *RNA*. The RNA is then translated into sequences of amino acids, and these sequences are fold up to form *proteins*. The proteins are then either used to build physical structures in the cell, or they catalyze specific chemical reactions in the cell. These specialized proteins are also called *enzymes*, and their task is to control the chemical processes in the cell. To wrap it up, the *biochemical* reactions are controlled by proteins, which are controlled by genes, giving rise to the Gene-Protein-Reaction (GPR) association.

Given the right genes, biochemical reactions can be streamlined to convert one chemical structure over several reaction steps into a totally different one, for example CO₂ into sugar, sugar into fat, and fat back to CO₂. This process is called *metabolism*: Every organism - whether it is a yeast cell in a beer fermenter, a tree or a lion - takes up nutrients (glucose vs. CO₂ and light vs. gazelle), extracts the energy from the input (energy of photons vs. chemical energy in glucose vs. gazelle meat), uses the energy and the chemical compounds to build and maintain itself and to move (yeast biomass vs. tree trunk vs. lion's fur), and finally excretes unused by-products (ethanol vs. oxygen vs. gazelle bones). Metabolism can therefore be described as the global integration of biochemical reactions, catalyzed by enzymes, which leads to the overall chemical conversion of energy and matter in a cell.

However, even though we can *read* the genetic code, we still do not fully understand how it is used to produce certain physiologies or to make organisms behave in a certain way.

There are two main reasons for this: The first problem is that linking a gene to a physiological output is not straightforward: The interactions between the many elements involved in a biological process give rise to complex behavior. In other words, even if we understand the exact function of each single gene, enzyme or metabolic reaction, the final biological outcome may be difficult to predict and even counter-intuitive. The second problem is that many processes and elements have not been characterized yet, or cannot be quantified easily in experiments. I will first discuss complex interactions and how to model them, and second examine the type of data that collected in biological research.

1.1.1 Systems biology

For decades, the reductionist approach has been the main driver for advancing biological. In this approach, single elements such as genes, proteins, organisms, diseases are studied in isolation. Reductionism has provided us with a detailed understanding of molecular mechanisms and their roles in nature, and lead to an accumulation of biological data and knowledge. However, it became more and more clear that understanding a single molecular process is not sufficient to explain a physiological output due to complex interactions between the different molecular actors. The problem of complex behavior in biology is addressed in systems biology⁴. In contrast to the reductionist approach, systems biology tries to obtain a holistic understanding of biological processes. Systems biology integrates acquired biological knowledge into mathematical models that can reproduce biological behavior⁵, or predict biological behavior under conditions that are not accessible experimentally (*e.g.*, the metabolism of intracellular parasites). These models translate the complex nature of a biological or biochemical process into a simplified mathematical description. A model consists of a set of mathematical equations that describe the process, and when fed with an input condition (*i.e.*, variables), will produce an output (*i.e.*, result values) using mathematical and computational techniques⁶. The model does not need to be an exact copy of the reality, but it should capture the main interactions of the process under study. Computational, or *in silico*, modeling approaches have been successfully applied to study cellular signaling cascades, metabolic fluxes, gene and protein interactions, cell-cell interactions, to name a few. Systems biology has entered many branches of biological research where it advances our understanding of complex phenomena such as cancer⁷, vector-borne diseases (*e.g.*, malaria), host-pathogen interactions, microbial communities, and it has potential applications in bioremediation⁸.

1.1.2 Data in biology: the “omics” era

The second problem that hampers the prediction of physiology from the genetic sequence is our incomplete knowledge of biological systems. In the past decade, technological advances in experimental techniques have flooded biology-related research fields with tremendous amounts of high-throughput data on different levels that allowed to quantify many elements at the same time. This new type of experimental acquisition of data is generally referred to as the “omics”. The omics approach regroups different biological quantification techniques which are organized by the type of biological elements they measure,

e.g., genomics, transcriptomics, proteomics, metabolomics, fluxomics. Technical advances in genome sequencing, for example, have led to an abundance of sequenced genomes of different organisms that are stored in publicly available databases. To determine which genes are expressed at a given moment in a cell, transcriptomics is used to quantify the available RNA or, in a similar way, proteomics can be applied to determine the presence of proteins in a cell. To capture the metabolic state of a cell, metabolomics techniques can quantify the chemical species present in the cell. The common feature of the omics approaches is that the cell's physiological state can be sampled at a given time point on different levels. These methods, however, generate huge amounts of data that need to be analyzed. Analyzing these data is challenging because of the complex, non-linear relationships between the sampled elements. Statistical analysis methods are usually used to analyze omics data, but they are not always sufficient to take full advantage of the highly informative data and to draw biological conclusions. Alternatively, the acquired data can be integrated into mathematical models that account for the complex relationships between the single elements measured in an -omics experiment. The synergy of systems biology and omics techniques can be used to refine biological models and to gain an overall understanding of cellular processes.

1.1.3 Synthetic biology and metabolic engineering

The systematic understanding of biological processes that has emerged in the past decades enabled scientists to rationally engineer biological systems. Engineering biological systems belongs to the field of synthetic biology, which has been promoted as a promising approach to help achieving the United Nations' Sustainable Development Goals⁹. In synthetic biology, existing biological parts (*e.g.*, transcription factors in gene regulatory circuits) are reassembled and modified to change the behavior of an organism and to design new biological functions¹⁰. Many tools in synthetic biology are based on genome editing, which has become easier and more accessible thanks to the development of CRISPR tools that enable precise editing of genomes. Synthetic biologists rely on well-established principles from other engineering disciplines, such as the Design-Build-Test cycle: A new process is designed on paper or computationally, implemented in biological system, and then tested for performance. If the engineering objective is not reached, the cycle will go into a next round of re-designing, building and testing. To go back to the original question: Can we harness the engineering principles from synthetic biology in combination with the modelling approach from systems biology and the large-scale data from the omics to engineer organisms to sustainably produce second-generation fuels, chemicals and pharmaceutical compounds¹¹?

To achieve this, we will have to re-engineer microbial metabolism using *metabolic engineering*. Microbes have been used to transform chemicals since the dawn of humanity in a well-known process called *fermentation*. Fermentation describes the conversion of certain molecules in food (*e.g.*, carbohydrates) into energy and other products (*e.g.*, ethanol) in the absence of oxygen. For example, the fungus *Saccharomyces cerevisiae*, commonly known as baking yeast, can convert complex sugars from cereals into alcohol. The fermentation of grape juice into wine is driven by the growth of a diversity of microbes like yeast, and lactic

acid bacteria are responsible for turning milk into yoghurt. With the advent of genetic engineering tools, non-native chemicals have been added to the repertoire of compounds produced biologically from simple sugars within microbial hosts organisms, such as second-generation biofuels (*e.g.*, bisabolene), pharmaceuticals (*e.g.*, artemisinin, opioids) and commodity chemicals (*e.g.*, 1,4-butanediol). A host organism, also sometimes called *chassis*, is an organism that is genetically engineered to produce a compound of interest. Usually, the host is a well-studied bacterium or a fungus for which molecular tools (*e.g.*, transformation vectors) are readily available, such as *Escherichia coli* or *Saccharomyces cerevisiae*. Current efforts are expanding the range of compounds that can be produced in microorganisms, and in parallel scaling up microbial production from laboratory to industrial scale. To achieve the scale up in terms of product range and production rate, it is essential to understand in detail the underlying metabolic processes.

1.2 Metabolism

Understanding metabolism to be able to engineer it for the production of fuels and chemicals is only one aspect of why studying it is important. First of all, metabolism is at the intersection between chemistry and biology, and understanding of how life works at the chemical level is a fundamental research question. Metabolism integrates genetic and environmental inputs, and the resulting metabolic state of a cell directly reflects its physiology. From a medical point of view, studying human metabolism is important to understand the mechanism of metabolic diseases and cancer as well as the influence of microbiota.

Metabolism can be organized in different parts with distinct functions: The core metabolism is responsible for the major mass and energy turnover, and can be divided in glycolysis and gluconeogenesis, pentose phosphate metabolism, tricarboxylic acid (TCA) cycle and pyruvate metabolism. The metabolites produced in the core can then be assimilated into amino acids, fatty acids, and other *biomass building blocks* (BBBs) that are used to build cellular structures such as proteins, lipids, or DNA. Studying the core metabolism is crucial to understand the overall energy and mass turnover, which is important in metabolic engineering to reroute core metabolic fluxes to optimize the production of a target compound. At the periphery of metabolism, secondary metabolic pathways produce compounds that are not directly necessary for the survival of an organism. Secondary metabolites constitute the chemical defense of plants and fungi, they are used as coloring agents by animals to enhance their reproductive success, or they are produced by plants to attract pollinators through attractive fragrances. In plants and fungi, secondary metabolism is particularly diverse, because unlike animals, these organisms cannot run away from their predators and need to rely on defensive chemical agents to protect them against grazing animals, parasites or diseases. Different organism might react very differently to these toxic compounds: For example, fragrant compounds produced by aromatic herbs as a chemical defense are appreciated by humans to spice food. Thanks to its high chemical diversity and its various effects on humans, secondary metabolism is an important source for therapeutic and recreational drugs.

The rate at which certain metabolites are consumed by an enzyme and turned into a product is called a metabolic flux. Metabolic fluxes are very dynamic and can change quickly as a consequence of environmental or genetic perturbations. The overall distribution of metabolic fluxes at a given time point is called the *fluxome*. Unlike the proteome, transcriptome or the metabolome, the fluxome cannot be measured directly: The distribution of metabolic fluxes has to be inferred from data points that are accessible experimentally, such as the rate of consumption of a substrate molecule, or the production of CO₂ by a cell culture. Hence, modeling approaches are indispensable to characterize metabolic fluxes.

1.2.1 Metabolic modeling

A systematic understanding of metabolism is fundamental for modeling metabolism. Metabolic models can help us characterize metabolic fluxes, predict the metabolic response of an organism, and identify knowledge gaps in our understanding of the organism. To model metabolism, we consider it as an open system where energy and mass can go in (*i.e.*, input) and flow out modified (*i.e.*, output). In between, metabolic reactions transform the input molecules into the output molecules through sequences of metabolic reactions, called pathways. The set of metabolic reactions present in an organism is called *metabolic network*.

One of the prerequisites for modeling metabolism is an exact representation of the cellular metabolic network. To obtain a full description of the metabolism of an organism, we can take advantage of the genome, and translate the encoded information through the GPR relationship into a set of reactions that can be performed in the cell. The network of all possible reactions, given the genetic sequence of an organism, is then used to construct a so-called *GEgenome-scale metabolic Model* (GEM). The GEM therefore describes all the biochemical reactions happening in a cell based on the genetic information. It is important that all the metabolites in the network are connected through reactions with each other to form a stoichiometrically correct, mathematical description of metabolism.

The mathematical framework can then be used to model the metabolic behavior of an organism in an approach called *constraint-based modeling*. For this, the stoichiometric constraints of the reactions are used to determine the possible solution space of metabolic fluxes in the cell. Additionally, flux constraints are added that represent the exchange of metabolites between the cell and its environment (*e.g.*, oxygen uptake, CO₂ secretion, ethanol production). The constraint-based model can then be used to optimize a specific goal (*e.g.*, maximization of growth, or production of target compound). For example, to simulate growth, a biomass reaction is added to the model that consumes the different BBBs that are used to build biomass in experimentally determined ratios. To maximize growth, we can now solve an optimization problem that maximizes the objective (*i.e.*, growth). Flux Balance Analysis (FBA) is performed to find the distribution of metabolic fluxes that optimizes the given objective¹². To explore the limits of the solution space described by the constraint-based model, we can perform Flux Variability Analysis (FVA), which finds the lower and upper bounds of all the fluxes in the model.

Usually, it is desirable to reduce the size of the solution space of the metabolic fluxes to increase the accuracy of biological conclusions drawn from the model. To this aim, different techniques have been developed to further constrain the solution space with additional data. For example, the model can be constrained by adding thermodynamic constraints and experimental data on metabolite concentrations in the cell and analyzed using Thermodynamic Flux Analysis (TFA)¹³. The solution space of a constrained-based metabolic model can be further constrained by integrating different kinds of data, such as transcriptomics and proteomics measurements or kinetic data, or by considering the production cost of enzymes^{14,15}.

Metabolic models are excellent frameworks to test the completeness of our knowledge about the biochemical processes in the cell. As long as a model cannot reliably reproduce the physiological properties of a cell, we can assume that we miss an essential aspect of the physiological phenomenon under study. One of the recurrent knowledge gaps in metabolic models are missing reactions in metabolic pathways that make the production of a biomass building block impossible, resulting in an infeasible model. This issue can be bypassed by adding artificial reactions that, by deduction, have to be in the organism to produce a given metabolite known to be produced by the cell. This process, called gap-filling, solves the problem in the model, but it leaves us with an orphan reaction that misses a catalyzing enzyme. The completeness of a model can be further tested by comparing model predictions with experimental data. For example, *in silico* knockout experiments remove reactions one by one from the model to predict the effect of gene knock-out experiments. In case the predictions do not match the experimental data, this means that the organism can somehow compensate the loss of the knocked-out step through an alternative pathway, and that our model misses this information. These examples illustrate that metabolic models are extremely useful to identify knowledge gaps and to direct future research efforts towards new discoveries.

1.2.2 Knowledge gaps in metabolism

The systems biology approach to metabolism not only taught us how single elements are linked in metabolism, but also gave us a systematic overview of missing links, or *knowledge gaps*. These gaps hinder efficient engineering of organisms – we need to fill the gaps and complete our knowledge. As discussed, metabolic models are extremely helpful to pinpoint blind spots in our knowledge of biochemistry. In the following, we will discuss four different types of knowledge gaps relevant in metabolism: (i) Unknown gene function, (ii) unknown reactions due to promiscuous enzymatic activity, (iii) orphan reactions, and (iv) orphan compounds (Figure 1.1).

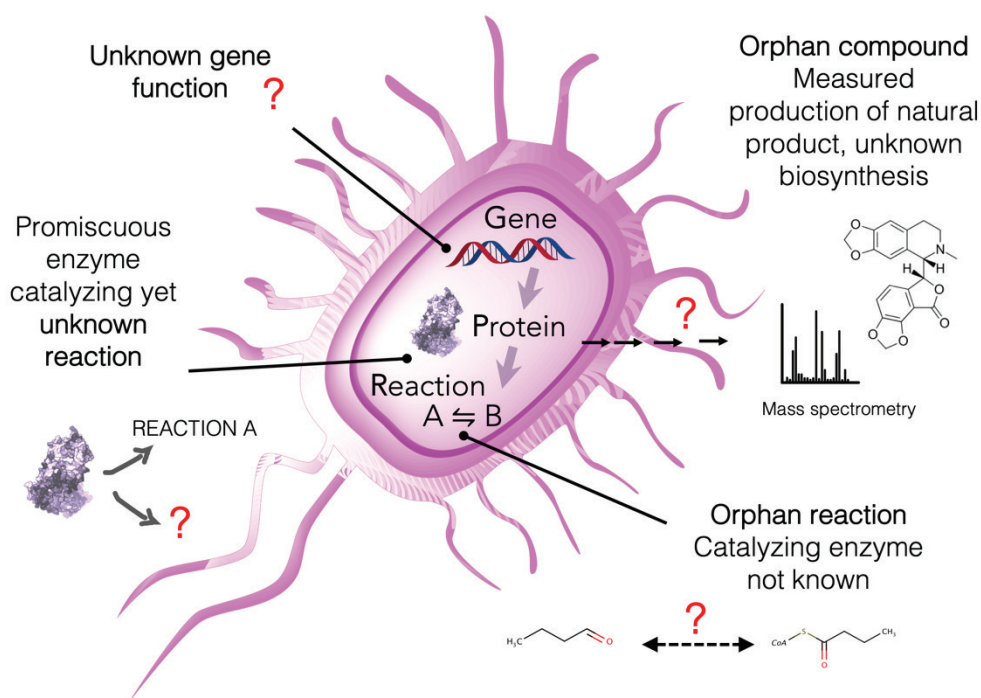


Figure 1.1: Different levels of unknowns in metabolism.

For many open reading frames (*i.e.*, sequences in the genome that can be translated into proteins, ORFs), the function of the encoded protein is not known. For example, 1,155 out of 4,505 (25%) of protein-coding genes in the well-studied model organism *Escherichia coli* are missing a functional annotation¹⁶. Since 1,567 out of the 3,350 annotated gene products are enzymes, we can assume that a bit less than half (~500) of the unknown gene products are enzymes. To assign functions to gene sequences, homology-based approaches are broadly used. For this, an amino acid sequence is compared to a database of manually annotated sequences: The more two sequences are similar, the higher the probability that the proteins perform the same function in the organism. If no similar sequence can be found, experimental approaches are required to determine the exact function of the gene in the organism.

Even if we know one function of a given protein sequence, it is possible that the protein also performs secondary functions. In metabolism, enzymes that catalyze secondary reactions that are different from their main, or native, catalytic activity are called *promiscuous*. Two types of promiscuous enzymatic activity are generally distinguished; *substrate* promiscuity, also called substrate ambiguity or multi-specificity, and *catalytic* promiscuity, also called moonlighting¹⁷. Substrate promiscuity means that one enzyme can catalyze the same type of reaction using different substrates, while the catalytic promiscuity is used to designate enzymes that can catalyze different types of biotransformations. When talking about enzyme promiscuity we usually mean substrate promiscuity, since this case is more common in nature. Since secondary catalytic activities are by definition weaker than the native function, they are more difficult to detect experimentally. Low catalytic side activities very often

remain undetected, unless the enzyme is tested specifically *in vitro* (or *in vivo*) for promiscuous enzymatic activity.

Another possible case is that a given enzymatic activity is known to exist, but no gene sequence has been assigned to it. These enzymes, whose activity has either been characterized *in vitro* or whose existence has been deduced from gaps in metabolic networks, are called *orphan enzymes*¹⁸. In 2013, 22% of EC numbers were found to be orphan¹⁹. This high percentage of unknown gene-protein-reaction assignments is problematic because it hampers the automatic reconstruction of GEMs by producing gaps in the metabolic networks that have to be filled manually. Moreover, it compromises our biochemical resources that are essential for the design of novel, industrially relevant bioproduction pathways.

Another aspect of missing knowledge is the existence of metabolites in nature for which no biosynthesis pathway is known. These metabolites have been identified through mass spectrometry experiments and are mostly derived from organisms with important secondary metabolisms, *i.e.* plants and fungi. In KEGG (Kyoto Encyclopedia of Genes and Genomes)^{20,21}, one of the most comprehensive databases for metabolic data, almost 10,000 biological compounds are not part of any metabolic reaction, meaning that their biosynthesis has not been characterized yet. These compounds are called *orphan compounds*, and many of them are secondary metabolites that are only produced in low abundance by specific organisms and therefore difficult to measure, or big, complex structures with long biosynthesis pathways that are complicated to characterize²¹.

The named knowledge gaps have been systematically addressed in the past. For example, since the problem of orphan enzymes was pointed out in 2004^{22,23}, the number of orphan enzymes could be decreased by community efforts from 38% to 22% within ten years. Other efforts systematically integrated orphan compounds in KEGG through bio- and cheminformatic approaches^{21,24}. Hence, if the problem can be named and quantified, it can be addressed more easily. However, it seems that what we know today is only a fraction of what possibly exists in nature. First of all, only known 1.3 million organisms have been described today, out of an estimated number of 8.7 million species²⁵. The number of sequenced organisms is even slightly lower, at less than one million according to the sequence database UniProt²⁶. From a chemical point of view, the number of known molecules in known organisms is , which is estimated to be only a small fraction of possible biochemical molecules, even when considering the tight constraints posed by biology on physical conditions such as temperature pressure and pH²⁷. These numbers strongly suggest that we have only explored a fraction of the chemo- and biodiversity present in nature, and that exploring and predicting the hypothetical biochemical space should be a priority to advance the field.

1.3 Metabolism at the level of atoms

Bridging the knowledge gaps in metabolism is crucial to complete our understanding of metabolic processes. To confirm the function of genes, explore promiscuous activities, find orphan enzymes and characterize biosynthesis pathways, the only approach that guarantees correct results is experimental. Unfortunately, experimental confirmation demands

high investments in terms of time, money, and workforce. Computational approaches are required to guide our efforts and to generate hypotheses about the true function of a gene, or the potential sequence of an orphan enzyme. Bioinformatics has provided efficient tools to assign function to genes by comparing gene sequences. In metabolism, modeling biochemical processes at the network level of reactions and metabolites has proven useful to find knowledge gaps, but these methods are not capable to provide hypotheses on how these gaps might be solved. For instance, a genome-scale metabolic network cannot predict the way how an enzyme rearranges atoms in a biochemical reaction, or how the activity of an enzyme changes as a consequence of a genetic modification. However, these aspects are particularly important if we want to predict the potential of enzymatic catalysis. We need cheminformatic tools to represent biochemical processes at a mechanistic level and to predict and explore the potential of enzymatic catalysis.

In this thesis, we address the need of computational approaches for the representation of enzymatic reaction mechanisms. We model enzymatic action *in silico* at the atomic level to generate hypothesis on the existence metabolic reactions in biological systems and to predict potential to-be-engineered biochemical functions.

1.3.1 Enzymes *in silico*

Several tools have been developed in the past to mimic the activity of enzymes using so-called enzymatic reaction rules. Generalized enzymatic reaction rules are used to predict biochemical reactions based on known enzymatic reaction mechanisms. An important feature of reaction rules is that they are “general”, meaning that one reaction rule always catalyzes a same type of biochemical reaction, but can apply this biotransformation on a range of substrates showing the same reactive site. Generalized reaction rules thus computationally mimic the promiscuous activity of an enzyme or a group of related enzymes. The rules can be derived manually by biochemists from biochemical knowledge^{28,29}, or they can be automatically extracted from known enzymatic reactions^{30–33}. Extracting reaction rules automatically from biochemical databases is fast, but the quality of the reaction rules will heavily depend on the correctness of the reference data. Furthermore, the automatic extraction of rules is prone to errors, especially in cases where the reaction mechanism is not easily derived from structural comparison between substrate and product. On the other side, manual derivation and curation of reaction rules is extremely time consuming, but it guarantees the biochemical correctness of the implemented reaction mechanisms.

For the named reasons, the work presented here is based on the computational reaction prediction tool BNICE^{34–36}. BNICE, which stands for Biochemical Network Integrated Computational Explorer, has at its core a comprehensive set of expert-curated, generalized enzymatic reaction rules. The development of BNICE has been started at the Northwestern University in 2005 as a tool for computational prediction of biochemical reaction network. Since then, BNICE has been applied to a diverse set of problems including the prediction of biosynthesis pathways for 2nd-generation fuels and commodity chemicals^{37,38}, the prediction of biodegradation pathways for xenobiotic compounds³⁹ and the exploration of

biosynthesis pathways of antibiotic polyketides⁴⁰ and lipids⁴¹. BNICE is also at the source of the ATLAS of Biochemistry database of novel reactions between known biological compounds⁴², and it has been used to construct the database MINEs that predicts potential metabolites derived from the known metabolome⁴³. The development has been continued independently at EPFL, and renamed into BNICE.ch. In the past ten years, its database of reaction rules has been continuously expanded from 291 in 2014 to 447 in 2019. In the process, many reaction rules have been created to cover enzymatic reactions from secondary metabolism with complex reaction mechanisms (*e.g.*, ring-forming (*S*)-norcoclaurine synthase or chalcone synthase) on the way toward a comprehensive modeling framework of biochemical reaction mechanisms.

BNICE.ch computationally describes metabolites, reactions and enzymes (Figure 1.2): In BNICE.ch, compounds are represented as Bond Electron Matrices (BEM), which describe the molecule as a mathematical graph where atoms are nodes, and bonds are edges. Enzymes are represented by reaction rules, which describe the possible reactive site configurations recognized by the enzyme, and the reaction mechanisms performed by the enzyme in the form of a matrix defining the bonds broken and formed during the reaction. Each reaction rule comes in a forward and a reverse version, which form together a *bidirectional* reaction rule. The BNICE.ch reaction rules are organized according to the Enzyme Commission (EC) classification, which assigns a four-digit classifier to each enzyme. The first level of the EC number defines the type biochemical reaction catalyzed by the enzyme (*i.e.*, transferases, hydrolases, lyases, isomerases, ligases and translocases). The second EC level defines the type of the functional groups that the enzymes act on, and the third level represents the cofactors involved in the reaction or another property used to further classify the enzyme. BNICE.ch reaction rules are defined up to the third EC level, which means that a given reaction rule can act on a range of different substrates harboring the defined reactive sites. In that sense, a reaction rule represents catalytic elasticity of an enzyme.

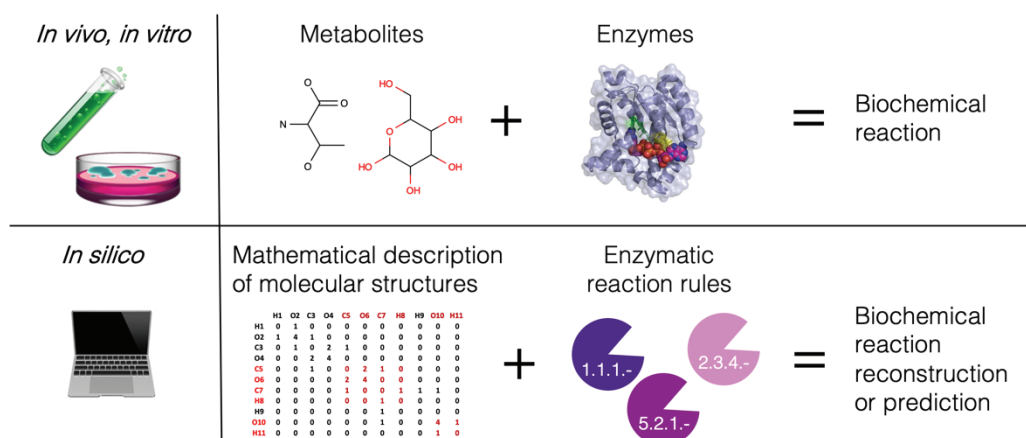


Figure 1.2: Computational representation of the biochemical actors in BNICE.ch.

The BNICE.ch reaction rules have three important characteristics: (i) The reaction mechanisms are encoded at the *atomic level*, and we can therefore use them to map atoms in biochemical reactions, (ii) since the exact substrate is not defined in the rule, we can apply them on a range of substrates harboring the same functional site to mimic *enzyme promiscuity*, and iii) we can take advantage of the description of the *reactive site* in the rule to find the reactive sites on a given substrate. In the following, we will discuss each aspect and its practical applications for biochemical *in silico* studies.

1.3.2 Atom-level resolution of reaction mechanism

The atom-level description of reaction mechanism in the BNICE.ch reaction rules form the methodological basis automatically to map atoms in biochemical reactions, and further to trace atoms through metabolic pathways and networks. By applying a reaction rule on a *in silico*-labeled substrate, BNICE.ch will rearrange the labeled atoms according to the reaction mechanisms stored in the rule and produce a labeled product, where every atom carries the label of its previous position in the substrate. To create labeled, linear pathways, we can to apply the next reaction rule on the labeled product, and to continue this procedure iteratively until the final metabolite of the pathway is produced. The manual derivation of the reaction mechanisms encoded in the reaction rules guarantees the correctness of the atom-mappings provided by BNICE.ch. Atom-maps are therefore in agreement with available biochemical knowledge, making them atom-maps particularly reliable. Applications of this feature of the BNICE.ch reaction rules are presented in Chapters 2 and 3 of this thesis.

1.3.3 Promiscuity-based prediction of biochemical reactions

A second important feature of the BNICE.ch reaction rules is the fact that only the functional group recognized by the enzyme is defined, but not the exact substrate. A single reaction rule can therefore recognize a broad range of substrates, and transform them into corresponding products according to the encoded reaction mechanism. This procedure not only reconstructs known metabolic reactions - it also predicts novel, hypothetical reaction that are feasible according to biochemical principles. These novel enzymatic activities are potentially performed by an enzyme in nature as a result of the promiscuous activity of the enzyme, or it can be engineered by genetically altering the binding pocket of the enzyme to allow the new substrate. The first case is of particular importance to fill gaps in metabolic networks and to discover yet uncharacterized biosynthesis pathways towards secondary metabolites (*e.g.*, plant natural products). Both cases form a valuable starting point for enzyme design in metabolic engineering, especially when the pathways to be engineered involve non-natural compounds. Chapter 4 of this thesis shows the value of reaction prediction in filling the gaps in our knowledge of metabolism, and Chapter 5 further illustrates the utility of enzyme prediction in different applications.

1.3.4 Predicting putative enzymes for novel reactions

The third crucial aspect of BNICE.ch reaction rules is its encoded definition of the reactive site that is recognized by the enzyme. For example, the encoded reactive site pattern can be used to quickly screen big numbers of molecules for potential biological activity. An application of this feature will be discussed in Chapter 4 (4.4.3). More importantly, the knowledge of reactive sites is key to predict putative enzymes for novel and orphan reactions: The in-house computational tool BridgIT takes advantage of the reactive sites encoded in the reaction rules to compare a query reaction to all known, enzyme-catalyzed reactions, and to calculate similarity scores between the reactions⁴⁴. Predicted enzymes can either catalyze the orphan reactions, or they can be genetically modified to accommodate the new substrate in their active site. BridgIT, related tools and their applications are discussed in detail in Chapter 5 (5.1.4).

1.4 This thesis

In this work, atom-level biochemical modeling is used to tackle different problems in the exploration, analysis and engineering of metabolism. We show that the computer-encoded knowledge of enzymatic reaction mechanisms has a broad range of applications, ranging from atom-mapping and -tracing in metabolism, over predicting novel, potentially existing or to-be-engineered biochemical reactions, to efficiently navigating big biochemical data in the search for metabolic pathways. The topics are presented in chapters as follows (Figure 1.3): In Chapter 2, we employ the mechanistic description of enzymatic action to obtain atom-maps for metabolic reactions, and to simulate stable-isotope tracer experiments *in silico*. Next, the experience gained from tracking atoms was exploited to develop a novel, efficient pathway search method that is capable to search large biochemical networks (Chapter 3). In Chapter 4, we use the promiscuous characteristic of generalized enzymatic reaction rules to systematically predict novel biochemical reactions around known biological and bioactive compounds, and we create a database series that hold the hypothetical biochemical networks and makes them accessible to the public. In Chapter 5, we present several applications of the aforementioned methods, including retrobiosynthetic pathway design to commodity chemicals and fuels, the prediction of pharmaceutical derivatives from known secondary metabolic pathways, and the prediction of biodegradation pathways for xenobiotic compounds. A final chapter summarizes the findings and proposes an outlook on future developments (Chapter 6).

Working in an interdisciplinary field like computational systems biology requires collaborating with other scientists specialized in specific tools or experimental techniques. During my PhD, I had the pleasure to work with many different people, and to ensure that they get the credits for their contributions, each Subchapter starts with a short statement on who has done what, printed in *grey italics*.

I wish you, dear reader, a pleasant time reading my thesis on *modeling, predicting and mining metabolism at atom-level resolution*.

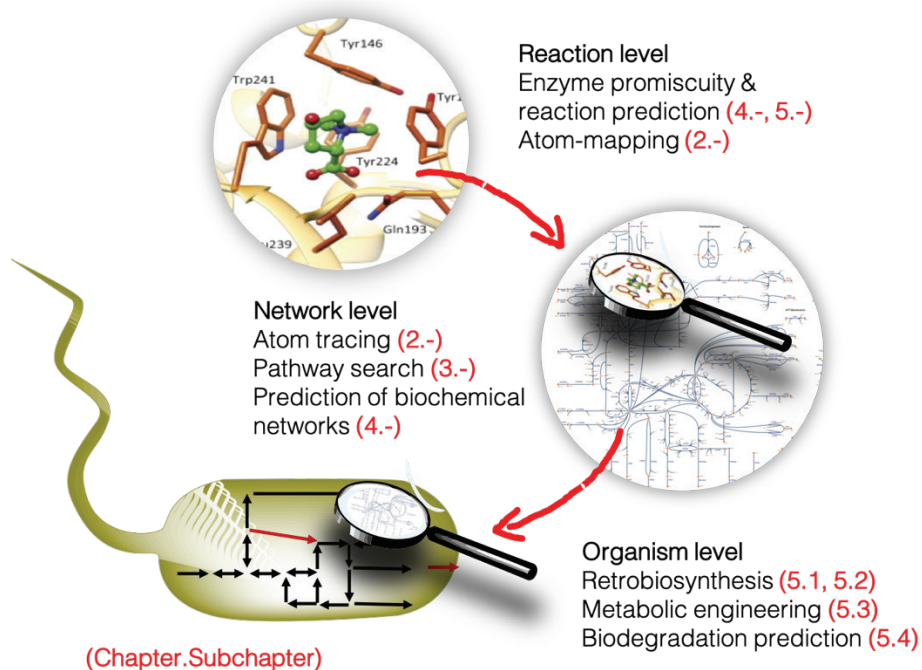


Figure 1.3: Overview on the different levels of metabolism addressed in this thesis. Red numbers indicate respective Chapters and Subchapters.

References

1. Arai, K., Smith, R. L. & Aida, T. M. Decentralized chemical processes with supercritical fluid technology for sustainable society. *J. Supercrit. Fluids* **47**, 628–636 (2009).
2. Linnaeus, C. *Systema Naturae. Holmiae* **1**, (Laurentii Salvii: Stockholm, 1758).
3. Darwin, C. On the origin of species. (1859).
4. Ideker, T., Galitski, T. & Hood, L. A NEW APPROACH TO DECODING LIFE : Systems Biology. *Annu. Rev. Genomics Hum. Genet.* **2**, 343–372 (2001).
5. Kitano, H. Computational systems biology. *Nature* **420**, 206–210 (2002).
6. Hillmer, R. A. Systems Biology for Biologists. *PLOS Pathog.* **11**, e1004786 (2015).
7. Zielinski, D. C. *et al.* Systems biology analysis of drivers underlying hallmarks of cancer cell metabolism. *Sci. Rep.* **7**, srep41241 (2017).
8. de Lorenzo, V. Systems biology approaches to bioremediation. *Curr. Opin. Biotechnol.* **19**, 579–589 (2008).
9. de Lorenzo, V. *et al.* The power of synthetic biology for bioproduction, remediation and pollution control: The UN's Sustainable Development Goals will inevitably require the application of molecular biology and biotechnology on a global scale. *EMBO Rep.* **19**, e45658 (2018).
10. Wang, Y.-H., Wei, K. Y. & Smolke, C. D. Synthetic Biology: Advancing the Design of Diverse Genetic Systems. *Annu. Rev. Chem. Biomol. Eng.* **4**, 69–102 (2013).
11. Choi, K. R. *et al.* Systems Metabolic Engineering Strategies: Integrating Systems and Synthetic Biology with Metabolic Engineering. *Trends Biotechnol.* (2019). doi:10.1016/J.TIBTECH.2019.01.003
12. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? *Nat. Biotechnol.* **28**, 245–8 (2010).
13. Henry, C. S., Broadbelt, L. J. & Hatzimanikatis, V. Thermodynamics-based metabolic flux analysis. *Biophys. J.* **92**, 1792–1805 (2007).
14. Pandey, V., Hernandez Gardiol, D., Chiappino Pepe, A. & Hatzimanikatis, V. TEX-FBA: A constraint-based method for integrating gene expression, thermodynamics, and metabolomics data into genome-scale metabolic models. *bioRxiv* 536235 (2019). doi:10.1101/536235
15. Salvy, P. & Hatzimanikatis, V. ETFL: A formulation for flux balance models accounting for expression, thermodynamics, and resource allocation constraints. *bioRxiv* 590992 (2019). doi:10.1101/590992
16. Keseler, I. M. *et al.* The EcoCyc database: reflecting new knowledge about *Escherichia coli* K-12. *Nucleic Acids Res.* **45**, D543–D550 (2017).

17. Khersonsky, O. & Tawfik, D. S. Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective. *Annu. Rev. Biochem.* **79**, 471–505 (2010).
18. Lespinet, O. Orphan Enzymes? *Science (80-.)*. **307**, 42a-42a (2005).
19. Sorokina, M., Stam, M., Médigue, C., Lespinet, O. & Vallenet, D. Profiling the orphan enzymes. *Biol. Direct* **9**, 10 (2014).
20. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
21. Kotera, M., McDonald, A. G., Boyce, S. & Tipton, K. F. Eliciting Possible Reaction Equations and Metabolic Pathways Involving Orphan Metabolites. *J. Chem. Inf. Model.* **48**, 2335–2349 (2008).
22. Karp, P. D. Call for an enzyme genomics initiative. *Genome Biology* **5**, (2004).
23. Roberts, R. J. Identifying protein function - A call for community action. *PLoS Biology* **2**, (2004).
24. Hadadi, N., Hafner, J., Shajkofci, A., Zisaki, A. & Hatzimanikatis, V. ATLAS of Biochemistry: A Repository of All Possible Biochemical Reactions for Synthetic Biology and Metabolic Engineering Studies. *ACS Synth. Biol.* (2016). doi:10.1021/acssynbio.6b00054
25. Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G. B. & Worm, B. How Many Species Are There on Earth and in the Ocean? *PLoS Biol.* **9**, e1001127 (2011).
26. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
27. Dobson, C. M. Chemical space and biology. *Nature* **432**, 824–828 (2004).
28. Hatzimanikatis, V. *et al.* Exploring the diversity of complex metabolic networks. *Bioinformatics* **21**, 1603–1609 (2005).
29. Wicker, J. *et al.* enviPath – The environmental contaminant biotransformation pathway resource. *Nucleic Acids Res.* gkv1229 (2015). doi:10.1093/nar/gkv1229
30. Delépine, B., Duigou, T., Carbonell, P. & Faulon, J.-L. RetroPath2.0: A retrosynthesis workflow for metabolic engineers. *Metab. Eng.* **45**, 158–170 (2018).
31. Campodonico, M. A., Andrews, B. A., Asenjo, J. A., Palsson, B. O. & Feist, A. M. Generation of an atlas for commodity chemical production in *Escherichia coli* and a novel pathway prediction algorithm, GEM-Path. *Metab. Eng.* **25**, 140–158 (2014).
32. Tyzack, J. D., Ribeiro, A. J. M., Borkakoti, N. & Thornton, J. M. Exploring Chemical Biosynthetic Design Space with Transform-MinER. *ACS Synth. Biol.* **8**, 2494–2506 (2019).
33. Sivakumar, T. V., Giri, V., Park, J. H., Kim, T. Y. & Bhaduri, A. ReactPRED: A tool to predict and analyze biochemical reactions. *Bioinformatics* btw491 (2016). doi:10.1093/bioinformatics/btw491

34. Hatzimanikatis, V. *et al.* Exploring the diversity of complex metabolic networks. *Bioinformatics* **21**, 1603–1609 (2005).
35. Soh, K. C. & Hatzimanikatis, V. DREAMS of metabolism. *Trends Biotechnol.* **28**, 501–508 (2010).
36. Hadadi, N. & Hatzimanikatis, V. Design of computational retrobiosynthesis tools for the design of de novo synthetic pathways. *Curr. Opin. Chem. Biol.* **28**, 99–104 (2015).
37. Henry, C. S., Broadbelt, L. J. & Hatzimanikatis, V. Discovery and analysis of novel metabolic pathways for the biosynthesis of industrial chemicals: 3-hydroxypropanoate. *Biotechnol. Bioeng.* **106**, 462–473 (2010).
38. Tokić, M. *et al.* Discovery and Evaluation of Biosynthetic Pathways for the Production of Five Methyl Ethyl Ketone Precursors. *ACS Synth. Biol.* acssynbio.8b00049 (2018). doi:10.1021/acssynbio.8b00049
39. Finley, S. D., Broadbelt, L. J. & Hatzimanikatis, V. Computational framework for predictive biodegradation. *Biotechnol. Bioeng.* **104**, 1086–1097 (2009).
40. Joanna González-Lergier, Linda J. Broadbelt, * and & Hatzimanikatis*, V. Theoretical Considerations and Computational Analysis of the Complexity in Polyketide Synthesis Pathways. (2005). doi:10.1021/JA051586Y
41. Hadadi, N. *et al.* A computational framework for integration of lipidomics data into metabolic pathways. *Metab. Eng.* **23**, 1–8 (2014).
42. Hadadi, N., Hafner, J., Shajkofci, A., Zisaki, A. & Hatzimanikatis, V. ATLAS of Biochemistry: A Repository of All Possible Biochemical Reactions for Synthetic Biology and Metabolic Engineering Studies. *ACS Synth. Biol.* **5**, 1155–1166 (2016).
43. Jeffryes, J. G. *et al.* MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. *J. Cheminform.* **7**, 44 (2015).
44. Hadadi, N., MohammadiPeyhani, H., Miskovic, L., Seijo, M. & Hatzimanikatis, V. Enzyme annotation for orphan and novel reactions using knowledge of substrate reactive sites. *Proc. Natl. Acad. Sci. U. S. A.* 201818877 (2019). doi:10.1073/pnas.1818877116

Chapter 2 Atom-level resolution of metabolic networks

The results presented in this Subchapter have been obtained in collaboration with several people under the project lead of the author of this thesis. The master student and later intern, Beatriz Lopes, contributed to the implementation of the iAM.NICE framework, applied it to E. coli and performed the substrate-utilization analyses on Plasmodium falciparum (Subchapter 2.4) under the supervision of the author. Furthermore, Dr. Anush Chiapino-Pepe implemented the ^{13}C -FBA framework, and Zhaleh Hosseini implemented and performed ^{13}C -FBA on E. coli. Subchapter 2.1 and 2.3 will be published as an article with the mentioned contributors as co-authors. Subchapter 2.2 introduces the previously published iAM.NICE workflow, developed by the author under the direct supervision of Dr. Noushin Hadadi.

Understanding metabolism at atom-level resolution is a difficult challenge, yet it is extremely useful to decipher the nature of metabolic processes. The following Chapter 2 is dedicated to modeling metabolism at atom-level resolution. The first Subchapter (2.1) introduces the importance of atom-level metabolic modeling and its importance to measure metabolic fluxes. Subchapter 2.2 discusses the atom-level representation of enzymatic reaction mechanisms as reaction rules, and how the knowledge encoded in reaction rules can be used to map atoms in metabolic reactions, pathways and networks. Next, we introduce a new approach to model stable-isotope experiments using ^{13}C -labeled glucose in Subchapter 2.3. Subchapter 2.4 shows an application of computational carbon tracing in the malaria parasite *Plasmodium falciparum*, and is followed by a conclusion (Subchapter 2.5).

2.1 Quantification of cellular metabolic fluxes

The distribution of metabolic fluxes defines the physiological state of the cell¹, and its study is crucial to advance our understanding of the cellular metabolic responses to environmental and genetic changes. With flux profiles, we quantify the flux percentages at each branching point in metabolism. This study of metabolic flux distributions, also called fluxomics, has been described as the functional output of the combined omics, *i.e.*, genomics, transcriptomics, proteomics and metabolomics². Hence, exact flux profiles are important to understand, analyze and compare metabolic processes in living organisms, and also to evaluate interactions between a cell's environment and its genetic material. In particular, we can further use the outcome of fluxomics studies in metabolic engineering efforts to design and optimize strains towards the creation of powerful cell factories. Currently, there are two popular methods to estimate metabolic fluxes inside cells: (i) stable-isotope tracing

experiments combined with ^{13}C -Metabolic flux analysis (^{13}C -MFA), and (ii) constraint-based methods seeking to optimize a biological objective, such as Flux Balance Analysis (FBA). In the following, the term “optimization-based methods” is used to refer to the latter, keeping in mind that ^{13}C -MFA also solves an optimization problem (*i.e.*, minimizing the difference between the experimental data and the underlying model).

2.1.1 ^{13}C -Metabolic flux analysis (^{13}C -MFA)

^{13}C -MFA combines experimental measurements with computational fitting to obtain metabolic flux distributions^{2–4}. ^{13}C -MFA relies on ^{13}C stable-isotope labeling experiments, where the carbon atoms of the substrate are partially or fully replaced with their stable isotope (^{13}C) (Figure 2.1). The isotope-labeled substrate is taken up by the cell, metabolized and incorporated into proteinogenic amino acids or other molecules that contribute to biomass. The distribution of isotope isomers (isotopomers) of different metabolites in the biomass is then measured experimentally, usually using gas chromatography-mass spectrometry (GC-MS). The experimental isotopomers data plus flux measurements for uptake and secretion rates are then used to infer flux distributions in the cell by solving a fitting problem. In order to correctly interpret the isotopomer data, atom-mapped metabolic networks are an important prerequisite. ^{13}C -MFA generally relies on small, hand-curated atom-mapped representations of the central carbon metabolism and of a selection of amino acid biosynthesis pathways.

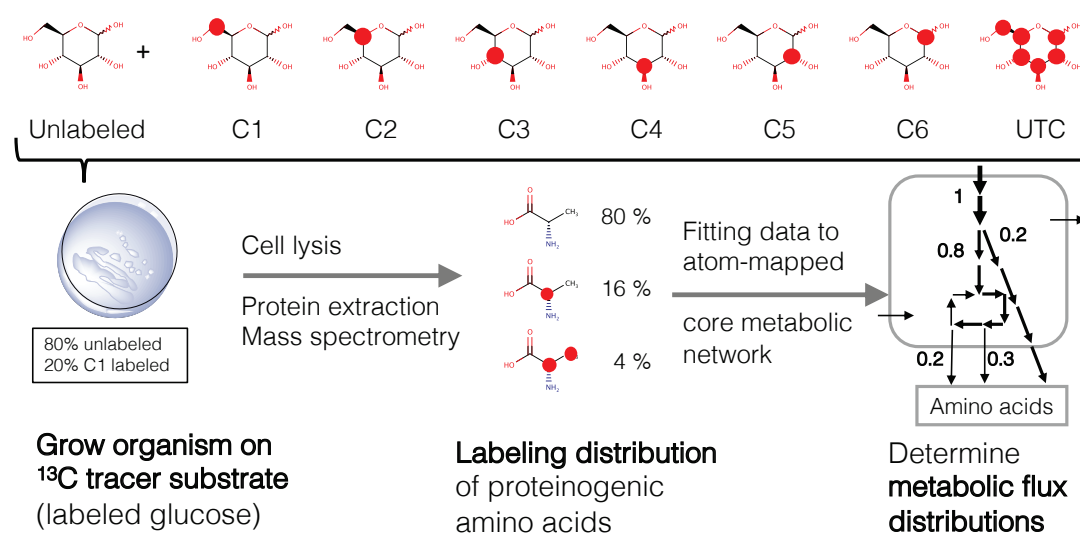


Figure 2.1: The ^{13}C -MFA workflow.

Please note that the carbon positions in glucose are usually numbered in the reverse sense than shown here. Hence, the isotopomer C1 is generally referred to as $[6\text{-}^{13}\text{C}]$, and C6 is $[1\text{-}^{13}\text{C}]$. UTC is also known as $[U\text{-}^{13}\text{C}]$.

However, these atom-mapped core networks have several shortcomings. First of all, there is no consensus in the set of represented metabolic pathways in the core network, and because of its small size spanning only the central carbon metabolism and amino acid

biosynthesis, the atom mapped networks fail to capture the totality of carbon transitions. The latter implies that overall stoichiometric constraints and full cofactors balances are usually not considered in atom-mapped core networks. Second, there is no systematic way to define the core model for non-model species. The standard core model only holds true for specific sets of species and cell cultures, thus compromising the determination of fluxes in other organisms and metabolic states⁵. This happens despite the fact that isotope labeling experiments have proven particularly useful to characterize the metabolism of non-model organisms, to identify bottlenecks and wasteful bioproduct pathways and to finally use the acquired knowledge for the metabolic engineering of those organisms⁶. Third, it has been shown that flux fitting within genome-scale models affects the flux ranges obtained by ¹³C-MFA, meaning that full metabolite and cofactor balance is crucial for obtaining reliable results⁷. Fourth, ¹³C-MFA requires solving a non-linear optimization problem, with fluxes as parameters, which may result in multiple locally optimal solutions for the flux distribution⁸. Fifth, *ad-hoc* assumptions of flux directionalities within the network may bias the results of the analysis. Finally, while confidence intervals for fluxes can be obtained from statistical analysis, the flux distribution obtained from ¹³C-MFA remains a unique solution of the flux state and does not embrace the more realistic concept of flux ranges as employed in optimization-based models.

The named shortcomings compromise our ability to fully take advantage of the information-heavy results from ¹³C labeling experiments, and therefore to accurately quantify the metabolic state of a cell. We believe that these issues could be solved by using organism-specific genome-scale models describing the overall stoichiometry of metabolism, followed by a robust, standardized reduction of the model and subsequent optimization-based flux analysis. However, it is currently not possible to directly integrate isotope-labeling data into this type of model.

2.1.2 Optimization-based methods to estimate flux ranges

FBA is a constraint-based modeling technique to calculate feasible flux ranges in a pseudo steady-state given the stoichiometric constraints of the reactions in the cell^{9,10}. FBA relies on the definition of an objective function (*e.g.*, growth) to describe the overall goal of the organism and the associated metabolic flux profiles. By optimizing the objective function, FBA calculates feasible flux distributions given the stoichiometric constraints. In order to narrow down the solution space, additional constraints such as growth rate and rates of metabolite exchange with the media (uptake and secretions) can be added. Thermodynamic Flux Analysis (TFA)¹¹ reduces the number of degrees of freedom in the system by computing the possible flux directionalities of each reaction based on feasible concentration profiles¹², and it further allows the integration of thermodynamic and metabolomics data.

FBA, TFA and other constraint-based modeling methods require stoichiometrically correct descriptions of metabolism, in the format of metabolic networks. In an ideal case, the metabolic network includes all the metabolic reactions that occur in a cell. The gold standard

for metabolic models are GEnome-scale metabolic Models (GEMs), which are directly derived from annotated genome sequences, and which comprehensively describe metabolic processes in the organism under study.

2.1.3 FBA *versus* MFA

In the analysis of labeling data, ^{13}C -MFA can directly evaluate metabolic profiles using precise quantification of stable-isotope species. However, the advantage of FBA vs ^{13}C -MFA lies in the usage of constraints derived from the integration of omics data sets to analyze alternative genome-wide metabolic profiles that describe a cellular state like the one measured with ^{13}C data. The complementarity of these two approaches has been nicely illustrated by Chen *et al.*, who used both FBA and ^{13}C -MFA in parallel to study internal metabolic flux distribution of *Escherichia coli* (*E. coli*)¹³. The flux distributions calculated by ^{13}C -MFA are frequently used to constrain metabolic models or to validate fluxes resulting from constraint-based modeling techniques. However, constraining a GEM with fluxes calculated from ^{13}C -MFA transfers biased assumptions of the small core model in ^{13}C -MFA to the GEM, a problem that has been discussed for a while^{2,14}. It has recently been addressed by Gopalakrishnan *et al.*, showing that adding proper stoichiometry, cofactor balance and biomass equations to the core network improves the accuracy of the flux distribution¹⁵. A similar approach has been proposed by García Martín *et al.*, where the fitting problem of ^{13}C -MFA was directly solved within the GEM in order to include cofactor balances and peripheral metabolic fluxes, such as the biosynthesis fluxes towards target compounds⁸.

As a conclusion, *integrating* labeling data directly into the FBA analysis would make it possible to take advantage of both the stoichiometric precision of the GEMs and the density of flux information inferred from stable-isotope experiments. Here, we suggest a new approach that, unlike previously published method, integrates isotopomer distributions in the form of constraints within FBA and TFA. With this approach, we systematically analyze metabolic flux profiles consistent with ^{13}C -labeling data for the study of metabolism and the production of all biomass building blocks (BBBs) at a genome-scale. We thus avoid the separate resolution of fitting problem for labeled fluxes and an optimization problem for the rest of the GEM.

The most important prerequisite to model isotopomer distributions in a cell is the ability to track single carbon atoms in a metabolic network. The problem of automatically obtaining biochemically correct atom-mapped reactions and tracing atoms through pathways and networks has been addressed in my master thesis and published in 2017¹⁶. The method, named iAM.NICE and summarized in the following Subchapter, forms the basis of constraint-based atom-level modeling of metabolism.

2.2 Tracing single atoms through reactions, pathways and networks

The iAM.NICE tool, published in Biotechnology journal, has been developed under the direct supervision of Dr. Noushin Hadadi (project lead, manuscript). Pipeline development, network generation and data analysis were done by the author of this thesis. This section presents the iAM.NICE workflow, as published in the Biotechnology Journal in 2016 by Hadadi et al.¹⁶.

Metabolic networks are typically drawn at the level of metabolites, which are connected by biochemical reactions. While this representation of metabolism is useful for studying metabolic processes at the network level, it does not provide any information on flow of atoms in metabolism. Yet, this information is crucial for the interpretation of stable-isotope labeling experiments to study the distribution of metabolic fluxes or the turnover of certain elements in metabolism. An atom-level representation of metabolic reactions is further crucial to understand the exact reaction mechanism performed by the enzyme.

Atom-maps can be derived by hand, although this approach is tedious. As an example, KEGG has discontinued the manual curated, semi-automatic approach of mapping atoms in reactions (RPAIR database¹⁷) due to the increased efforts of curating a fast-growing biochemical database. As an alternative, computational approaches can be used to map atoms in chemical and biochemical reactions. There are two main approaches to the problem based on graph theory^{18,19}: The first approach seeks to optimize the conservation of molecular substructures by looking for the Maximum Common Subgraph (MCS) between the reactants and the products^{20–26}, and the second approach tries to minimize the number of bonds broken and formed during the chemical transformation by determining the minimal chemical distance^{27–31}. Finally, mixed approaches include training data have been proposed in the past to decrease the number of erroneous mappings³². Mixed approaches are also available as AutoMapper in ChemAxon and as Atom Atom Mapping Tool in the Reaction Decoder Tool (RDT). However, none of the mathematical approaches can guarantee the correctness of its atom mappings; The underlying problem is that the mathematical solution is not necessarily the biochemically correct one, and the true atom map of a reaction may only be determined by studying the enzymatic reaction experimentally using labeled substrates. For example, mapping atoms in biochemical reaction catalyzed by ligases and isomerases are particularly challenging because these enzymes can perform complex rearrangement of atoms that are not captured by automatic atom mapping tools. Hence, in order to know the true atom mapping of a biochemical reaction, one needs to know the reaction mechanism performed by the catalyzing enzyme. In the following, we present a method for automatic atom mapping of reactions on the basis of generalized enzymatic reaction rules that encode the atomic rearrangement performed by an enzyme. The reaction rules are expert-curated representations of biochemical knowledge, thus guaranteeing the biochemical correctness of the resulting atom maps. The algorithm is called iAM.NICE and can be employed to map atoms in reactions and pathways, and it can finally be used to track atoms through complex metabolic networks.

2.2.1 Atom-mapped biochemical reactions

iAM.NICE is an extension of BNICE.ch, and it stands for *in silico* Atom Mapped Network Integrated Computational Explorer. At its core are the BNICE.ch enzymatic reaction rules, which describe the exact mechanism of an enzymatic reaction at the atomic level (Figure 2.2): The substrate, represented by a bond-electron matrix (*substrate matrix*), is recognized by the active site of an enzyme, represented by the *recognition matrix* in the reaction rule. The reaction mechanism, encoded in the *operator matrix*, is then applied to the part of the substrate recognized by the reaction rule. The result is a *product matrix* that describes the molecular structure of the product molecules. In order to generate an atom map of the reaction, we label the atom positions in the substrate and we apply the reaction rule that reconstructs the reaction we want to analyze. The reaction rule then transfers the label from the atom position in the substrate to the corresponding atom position in the product, thus creating an atom-map of the reaction.

Exploiting the reaction mechanism encoded in the expert-curated reaction rules to map atoms ensures that the atom mapping is biochemically correct. Given the biochemical correctness of the encoded reaction mechanisms, the reaction rules can further be used to resolve unclear mechanisms in metabolic reactions

Another advantage of this approach is that we can now go one step further and apply a consecutive reaction rule on the labeled product of the first reaction, thus mapping atoms over two or more sequential reaction steps. In the following, we will focus on tracing carbon atoms throughout reactions, pathways and networks. It should be noted, however, that the method can readily be applied to study the metabolic fate of elements other than carbon.

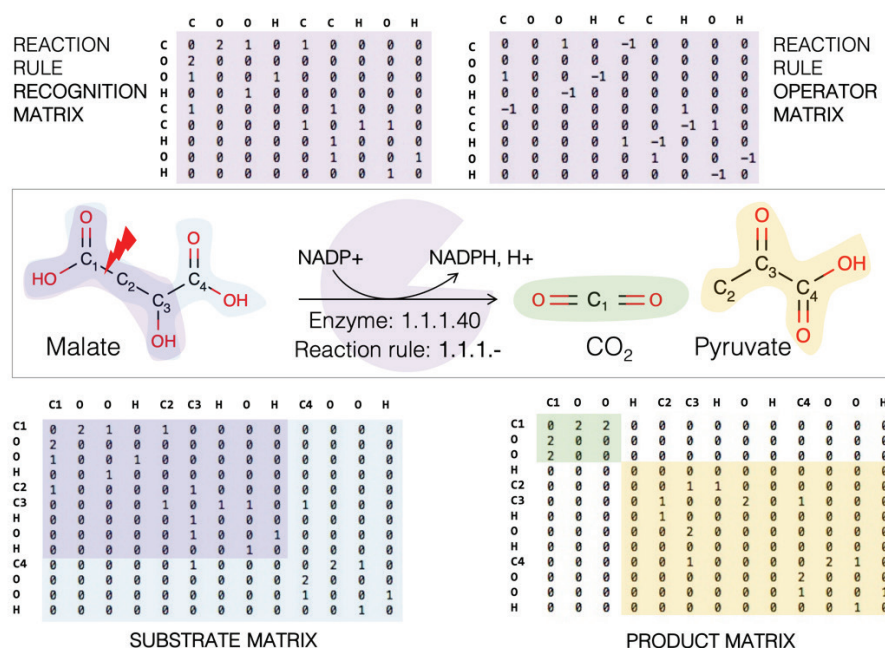


Figure 2.2: Generalized reaction rules represent the action of a substrate-promiscuous enzyme at atom-level resolution.

2.2.2 Atom-mapped metabolic pathways

Being able to trace each atom throughout a metabolic pathway is crucial to evaluate, for example, the carbon efficiency of a biosynthetic pathway in metabolic engineering, and to interpret the outcome of stable-isotope labeling experiments. By applying BNICE.ch reaction rules iteratively on a substrate, we can map atoms through metabolic pathways (Figure 2.3). In a first generation, the reaction rule representing the first reaction in the pathway is applied to the initial substrate, generating a labeled product. In a second generation, the reaction rule representing for the second biotransformation is applied the product of the first reaction, and so on, until the final, labeled product of the pathway is produced. The result is an exact mapping of each atom position in the precursor(s) of the pathway to the product molecule(s).

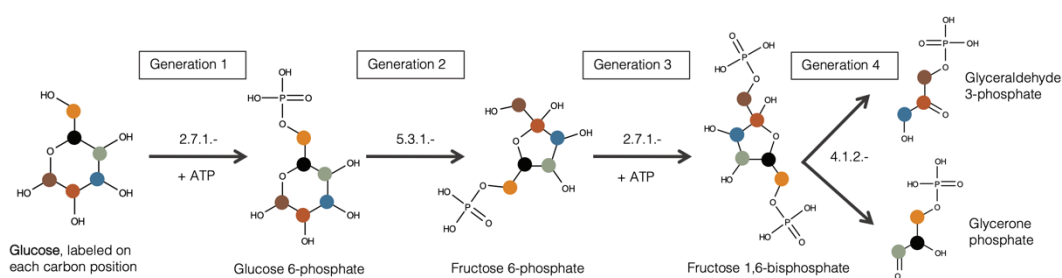


Figure 2.3: All the carbon atoms of glucose are traced simultaneously through the first four reaction steps of glycolysis in four generations.

2.2.3 Tracing atoms in the substrate through metabolic networks

Tracing single atom position in a substrate throughout a whole metabolic network can help us to determine the flow of elements in metabolism or to study the usage of molecular moieties within the metabolic network. Tracing atoms *in silico* through metabolic networks has also been used in the past to find metabolic pathways^{25,33–35}. For linear pathways, mapping all the atoms of the substrate simultaneously to the final product is straight-forward. However, as soon as certain metabolites are recycled in the pathway, as it is the case in metabolic networks, the emerging cycles will contribute to the scrambling of the labels. In these cases, it is more informative to look at a single atom at a time to reduce the complexity of the resulting atom map. To trace single carbon atoms from a substrate through a metabolic network, the same procedure is used as for linear pathways; Reaction rules are applied iteratively on a *in silico* labeled substrate until no more isotopomers (*i.e.*, labeled metabolites) are generated. The result is a map of all possible paths that the labeled atom can take within the reaction constraints of the network. For example, the carbon at position four in glucose can end up in two different positions in the downstream metabolite pyruvate, as illustrated in Figure 2.4. It is important to note that even though we trace only one single atom at time, the overall atom map of the network is still conserved in the reaction rules. This tracking of atomic positions *in silico* is equivalent to ¹³C-labeling experiments,

2.3 Atom-level modeling of *E. coli*

Now that we know how to trace atoms through metabolic networks, the next aim is to integrate the carbon-tracing network into constraint-based modeling framework. Here, we present a workflow for systematic reconstruction and analysis of atom-mapped genome-scale models, termed ^{13}C -FBA and illustrated in Figure 2.5. ^{13}C CFBA addresses the limitations of previously developed ^{13}C -MFA and FBA frameworks, as previously discussed in Subchapter 2.1. The aim of our work is to achieve the integration of atom-level information into a constraint-based modeling approach in order to allow direct integration of stable-isotope labeling data. For this, we needed a well-studied model organism for which labeling data is available, as well as a GEM of the organism. In a first step, we organize and reduce the GEM. Second, we apply thermodynamic network analysis to determine the directionalities of the reactions in the GEM. Third, we annotate the reactions with reaction mechanisms (*i.e.*, BNICE.ch reaction rules), which are used by iAM.NICE in step four to track the labeled atoms of different glucose tracers throughout the network. In step five, the labeled networks are then merged into the constraint-based framework, which is used to predict labeling patterns in the biomass building blocks. The final, hybrid model is finally used to study the propagation of the labels through the model and for the integration of experimental ^{13}C labeling data. The following sections describe each step of the workflow in detail. The workflow is illustrated by the application of ^{13}C -FBA on the GEM of *E. coli*, and we finally compare the outcome with the results obtained by similar approaches.

2.3.1 Reduction of the *E. coli* GEM

To illustrate our workflow, we chose *E. coli* as a model organism for its well-curated metabolic models, as well as for the abundance of available scientific literature and experimental labeling data. Here, we used the GEM created by Orth *et al.*, iJO1366³⁶. The first step of our workflow consists of reducing the GEM of *E. coli*. To achieve this, we applied a bottom-up systematic method proposed by Ataman *et. al*, called redGEM, which reduces the complexity of a genome-scale model into a core model³⁷. The complementary method of redGEM, called lumpGEM³⁸, finds biosynthesis pathways towards the BBBs and lumps them into single reactions that represent the stoichiometry of the whole lumped pathway. The reduced model preserves key properties from the original model such as biomass production, by-product yield, concentration variability and gene essentiality³⁷. Hence, it matches the requirements discussed previously of a consensus model suitable for atom-level reconstruction. We applied the systematic reduction on the *E. coli* GEM to generated a reduced metabolic model, called redEcoli, that serves as a reference for subsequent atom-mapping studies. The original model counts 1805 metabolites, 2583 reactions and 102 BBBs. redEcoli counts 309 metabolites and 539 reactions, which are composed of 119 metabolic reactions, 188 lumped reactions, 231 transport reactions between the cytosol, the periplasm and the extracellular space, and one biomass reaction. The 119 metabolic reactions that are part of the selected core subsystems glycolysis/gluconeogenesis, pentose phosphate pathway, TCA cycle, pyruvate metabolism, and the electron transport chain. The 102 BBBs are produced from the core metabolism via the 188 lumped reactions.

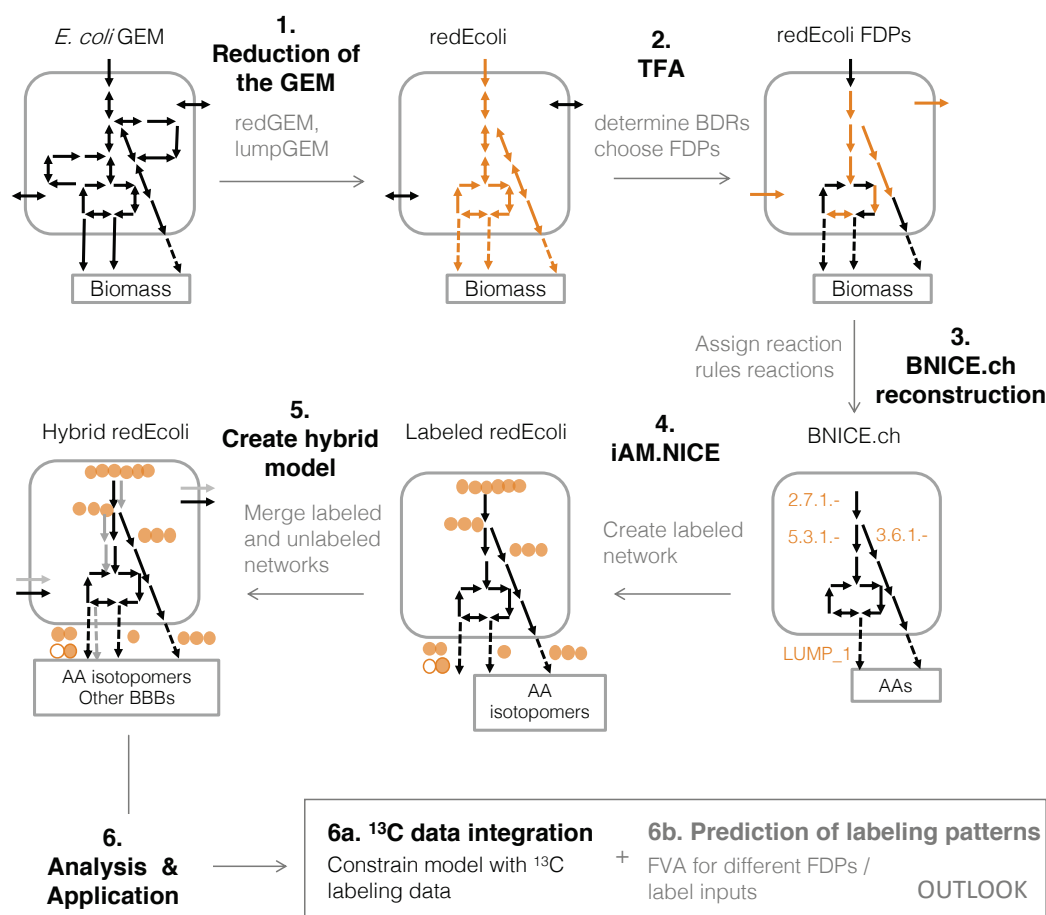


Figure 2.5: The ^{13}C -FBA workflow explains the creation of a reduced hybrid model in five steps, plus a final step for analysis and model validation. Abbreviations: genome-scale model (GEM), Thermodynamic Flux Analysis (TFA), flux directionality profile (FDP), bidirectional reactions (BDRs), amino acids (AA), biomass building blocks (BBBs). The third-level EC numbers in step three represent enzymatic reaction rules, and LUMP_1 stands for a lumped reaction rule converting core precursors into BBBs.

2.3.2 Thermodynamic analysis of redEcoli

In the second step of the workflow, we aim to further constrain redEcoli by determining the thermodynamically feasible flux directionalities for each reaction in the model in aerobic conditions, assuming maximal cell growth. For this, we curated redEcoli thermodynamically following the standard approach defined within the TFA framework^{39,40}. We then performed a Thermodynamic Variability Analysis (TVA) of redEcoli to identify the directionalities of the reactions that satisfy mass balances of all metabolites and cofactors, as well as thermodynamic constraints at predefined intracellular conditions (see Materials and Methods) to support growth. Nine reactions were found to operate both in the forward and reverse direction, which we call bi-directional reactions (BDRs) in the following. These reactions are catalyzed by the enzymes acetaldehyde dehydrogenase (ACALD), fumarase (FUM), malate dehydrogenase (MDH), phosphoglucose isomerase (PGI), ribulose 5-phosphate 3-epimerase (RPE), transaldolase (TALA) and transketolase 1 and 2 (TKT1, TKT2), and triose-

phosphate isomerase (TPI). A theoretical number of 2^9 (512) flux directionality profiles (FDPs) would result from all possible combinations of directionalities of the nine reactions. From those, we selected six FDPs supporting maximal growth and representing changing flux directionalities in different metabolic subsystems. The six selected FDPs affect the flux directionalities of the four enzymes ACALD, FUM, PGI and TALA, representing changes in the pyruvate metabolism, citric acid cycle, glycolysis and pentose phosphate pathway, respectively (Table 2.1).

Table 2.1: Six different FDPs were chosen for further analysis, considering four (green rows) out of nine bidirectional reactions (BDRs). 1 indicates that the reaction is operating in the forward direction in the specific FDP, and -1 in the reverse direction.

BDRs	FDP1	FDP2	FDP3	FDP4	FDP5	FDP6	Reaction equation
ACALD	-1	-1	1	1	1	1	Acetaldehyde + CoA + NAD ⁺ \rightleftharpoons Acetyl-CoA + H ⁺ + NADH
FUM	1	1	1	1	-1	1	Fumarate + H ₂ O \rightleftharpoons L-Malate
MDH	1	1	1	1	1	1	L-Malate + NAD ⁺ \rightleftharpoons Oxaloacetate + H ⁺ + NADH
PGI	1	1	1	1	-1	-1	D-Glucose 6-phosphate \rightleftharpoons F6P
RPE	1	1	1	1	1	1	D-Ribulose 5-phosphate \rightleftharpoons Xu5P
TALA	-1	1	-1	1	1	1	G3P + S7P \rightleftharpoons E4P + F6P
TKT1	1	1	1	1	1	1	R5PP + Xu5P \rightleftharpoons G3P + S7P
TKT2	1	1	1	1	1	1	E4P + Xu5P \rightleftharpoons F6P + G3P
TPI	1	1	1	1	1	1	Dihydroxyacetone phosphate \rightleftharpoons G3P

G3P: Glyceraldehyde 3-phosphate, S7P: Sedoheptulose 7-phosphate, E4P: D-Erythrose 4-phosphate, Xu5P: D-Xylulose 5-phosphate, F6P: D-Fructose 6-phosphate, R5PP: Alpha-D-Ribose 5-phosphate

2.3.3 Curation of redEcoli with reaction mechanisms

In a third step, we obtained an atom-mapping model of redEcoli by applying iAM.NICE, which, to our knowledge, is the only computational tool for automatic mapping of single atoms in metabolic reactions, pathways and networks that ensures the correctness of the mapping based on biochemical reaction mechanisms. iAM.NICE is an extension of the retro-biosynthesis tool BNICE.ch⁴¹, which has initially been developed to predict biochemical reaction networks using expert curated, generalized biochemical reaction rules. Each reaction rule encodes the exact reaction mechanism of an enzyme in a general way, meaning that the rule can apply its biochemical transformation on a set of substrates harboring the same reactive site. A single reaction rule can be applied on a range of *in silico* labeled substrates to generate biochemically correct atom maps for the resulting metabolic reactions.

To curate the reactions in redEcoli with reaction mechanisms, we first categorized the 119 metabolic reactions in the core of the reduced model. They included 22 reactions that only involved cofactors that were not produced by the core and hence would not get labeled from glucose in the core, as well as four polymerization reaction reactions that were excluded from the reaction rule assignment, which left us with 93 core reactions for reaction rule assignment. 65 generalized reaction rules were required to represent all the reaction mechanism in redEcoli. The result of this procedure is an atom-mapping model of the *E. coli*









core metabolism, annotated with reliable, biochemically confirmed reaction mechanism that can be readily used to map and track atoms.

2.3.4 Atom-mapped core metabolism of redEcoli

Once an atom-mapped model of the core has been established, we can feed it with glucose molecules labeled in different carbon positions and study the distribution of the *in silico* labels throughout the network. To analyze the atom-level model of the *E. coli* core metabolism, we first investigated the effect of different physiological states, represented by Flux Directionality Profiles (FDPs), and labeling positions in glucose on the generation of carbon-labeled network. In a second labeling experiment, we labeled the BBBs differently in each carbon position, and we traced back the possible origins of each atom in the precursor metabolites of the core. Next, we traced carbons through the whole redEcoli, from the substrate to the BBBs, in order to create carbon-mapped networks for a range of specific labeling configurations of the tracer substrate glucose, that could finally be analyzed using FBA-related tools.

Using iAM.NICE, we generated 28 core labeled networks that were a product of the combination of six singly-labeled glucose molecules with the six named FDPs (Table 2.2). The 28 labeled models are of different size regarding the number of labeled reactions and metabolites involved, which depends on the position of the labeled carbon in glucose and the FDP applied for the network reconstruction. The lack of assumed directionality in the nine bi-directional reactions is referred to as noFDP. Since noFDP has less directionality constraints, more combinations of isotopomers are possible which leads to bigger isotopomer networks than the FDPs 1 to 6. Furthermore, the number of produced isotopomers is highest for carbon label positions 4, 5 and 6 in glucose and smallest for carbon label at position 2. The theoretical maximum of isotopomers (1435) is only obtained in the case of unconstrained bi-directional reactions in noFDP, combined with UTC, C5 and C6 glucose. The FDPs 5 and 6 are identical in their numbers of generated isotopomers, indicating that the reaction directionality of FUM in the TCA cycle does not impact the labeling distribution. The opposite is the case for glycolysis enzyme PGI, which, if operating in the reverse direction (FDPs 5 and 6), more than triples the number of isotopomers for C4 and C5, and reduces by three the number of isotopomers for C6. The difference between the FDPs gives a first hint on the different flux distribution between different physiological states of a cell, and it highlights the importance of pre-assumed flux directionalities in the generation of atom-mapped networks.

Table 2.2: Number of compounds for each combination of an FDP with a labeled glucose substrate. Different isotopomers are counted as different compounds.

								
FDP	Unlabeled	C1	C2	C3	C4	C5	C6	UTC
1	68	377	137	94	417	404	381	417
2	68	377	137	94	429	438	383	449
3	68	371	135	94	411	398	375	411
4	68	371	135	94	423	432	377	443
5	68	371	135	94	1429	1429	81	1429
6	68	371	135	94	1429	1429	81	1429
none	68	402	141	402	1435	1435	1435	1435

2.3.5 Atom-mapped biomass production pathways

Next, we had to formulate lumped reaction rules that would map the atoms in the precursors to the atoms in the BBBs, in order to reconstruct the lumped reactions obtained from lumpGEM. For this, we first analyzed a total of 70 biosynthesis subnetworks for twenty BBBs. lumpGEM generated four levels of subnetworks: S_{min} is the set subnetworks that connect the precursors to a given BBB using the least number of reaction steps. $S_{min}+1$ is the set second-shortest subnetworks, and so on, until $S_{min}+3$. Using iAM.NICE in reverse mode, we traced the carbon atoms of 20 BBBs in *E. coli* through the biosynthesis subnetworks back to its core precursors using iAM.NICE. We analyzed the resulting carbon-mapped subnetworks from core precursors to BBBs for a total of 30 subnetworks (Table 2.3), which were found to be representative for all of the 70 initial subnetworks. The reason for this is that many subnetworks only differed in their usage of cofactors, which did not affect the carbon labeling. We finally used generated carbon maps to formulate lumped reaction rules.

Table 2.3: The subnetworks towards 20 BBBs have been generated by lumpGEM. Grey shades indicate representative subnetworks that are used to create lumped reaction rules and that are visualized online. The full information on subnetworks generated by lumpGEM and their reconstruction in iAM.NICE can be found in the Appendix (Table A1).

Biomass Building Block	Name of subnetwork	Degree	Number of reactions	Category
Proteinogenic amino acids				
Alanine	S1_ala	Smin	1	B
		Sminp1	2	
Arginine	S1_arg	Smin	9	B
		Sminp1	11	
Asparagine	S1_asn	Smin	1	B
		Sminp1	2	
Cysteine	S1_cys	Smin	11	B
		Smin	11	
		Sminp1	13	
		Sminp1	13	
Glutamine	S1_gln	Smin	1	A
Glycine	S1_gly	Smin	6	C
		Sminp1	7	
		Sminp1	7	
		Sminp2	8	
Histidine	S1_his	Smin	17	B
		Smin	17	
		Sminp1	18	
		Sminp1	18	
		Sminp2	19	
		Sminp2	19	
		Sminp3	20	
		Sminp3	20	
Isoleucine	S1_ile	Smin	10	A
Leucine	S1_leu	Smin	9	A
Lysine	S1_lys	Smin	9	A
Non-amino acid building blocks				
Chorismate	S1_chor	Smin	7	A
Putrescine	S1_ptrc	Smin	6	A

Biomass Building Block	Name of subnetwork	Degree	Number of reactions	Category
Proteinogenic amino acids				
Methionine	S1_met	Smin	21	C
		Smin	20	
	S2_met	Smin	20	
		Smin	21	
	S3_met	Sminp1	22	
		Sminp1	22	
	S4_met	Sminp1	22	
		Sminp1	22	
	S5_met	Sminp1	22	
		Sminp1	22	
	S6_met	Sminp1	22	
		Sminp1	22	
	S7_met	Sminp2	23	
		Sminp2	23	
	S8_met	Sminp2	23	
		Sminp2	23	
	S9_met	Sminp2	23	
		Sminp2	23	
	S10_met	Sminp3	24	
		Sminp3	24	
	S11_met	Sminp3	24	
		Sminp3	24	
Phenylalanine	S1_phe	Smin	10	A
Proline	S1_pro	Smin	4	B
		Sminp1	6	
Serine	S1_ser	Smin	3	A
Threonine	S1_thr	Smin	5	A
Tryptophan	S1_trp	Smin	15	C
		Sminp2	17	
		Sminp3	18	
Tyrosine	S1_tyr	Smin	10	A
Valine	S1_val	Smin	4	B
		Sminp1	5	

To validate the labeled subnetworks, we compared them manually to the lumped reactions used in traditional ^{13}C -MFA. As a reference, we use a state-of-the-art atom-mapping model for ^{13}C -MFA published by Leighty, R. W. & Antoniewicz in 2013⁴². The reference model has some systematic differences with our model. For instance, the reference model does not explicitly write the oxidized form NAD(P)^+ , and H_2O and H^+ molecules are omitted. Furthermore, our lumped reactions always describe the biosynthesis starting with core metabolites, while the reference model uses BBBs as precursors for the biosynthesis of downstream amino acids. *E.g.*, we start the biosynthesis of isoleucine from aspartate, while the reference model starts from the intermediate threonine.

The lumped reactions for the 20 amino acids under study can be classified into three categories: (A) single lumped reactions, (B) multiple lumped reactions with conserved carbon transformation, and (C) multiple lumped reactions with different carbon transformation (Table 2.3).

- A. Single lumped reactions (including chorismate, glutamine, isoleucine, leucine, lysine, phenylalanine, putrescine, serine, threonine, tyrosine): The atom maps for these lumped reactions are exactly the same as in the reference model, except that chorismate and putrescine are not present in the reference model.
- B. Multiple lumped reactions with conserved carbon transformation (including alanine, arginine, asparagine, cysteine, histidine, proline). For these cases, alternative lumped reactions use different cofactors, but the carbon transformation remains unchanged. The lumped reactions for these amino acids are exactly the same as the ones used in the reference model, with the exception of the histidine biosynthesis pathway, where the cofactor usage is slightly different; Our model uses glutamate and ammonia where the reference model uses glutamine as a nitrogen source, and we produce formate where the reference model produces 10-formyltetrahydrofolate. Furthermore, different subnetworks consume different numbers of ATP molecules, varying from four to six. The reference model consistently uses five ATPs.
- C. Multiple lumped reaction with different carbon transformation (including glycine, methionine and tryptophan): For these three amino acids, we found differences in terms of labeling between the alternative subnetworks.

The biosynthesis pathway of glycine starts from the core metabolites aspartate and 3-phosphoglycerate. In the case of aspartate, two carbon atoms are freed to end up in either acetaldehyde or formate. The single carbon atom liberated from 3-phosphoglycerate forms formate. In the reference model, the two precursors are the same, one producing glycine from aspartate through threonine, and the other producing glycine from 3-phosphoglycerate via serine. The difference lies in the carbon byproduct: the first pathway (via aspartate) produces acetaldehyde or acetyl-CoA, while the reference model only produces acetyl-CoA. The second pathway (via serine) sends the carbon to formate, while the reference model produces 5,10-methenyltetrahydrofolate. These differences in glycine production between our model and the reference model mainly result from the different definitions of the scope of the lumped reactions, and do not necessarily reflect a biochemical disagreement.

The second amino acid with carbon-changing biosynthesis pathways is methionine. The main backbone of the methionine molecule consists of four carbon atoms coming from the precursor aspartate (Figure 2.6A). The S-methylation is catalyzed by methionine synthase (METS), which takes the methyl from the folate one-carbon pool. According to the genome-scale model of *E. coli*, this carbon pool can be supplied by either formate, 3-phosphoglycerate or aspartate. In the reference model, the methyl group comes directly from a predefined folate one-carbon subnetwork. Since our model does not have a separate pool for one-carbon metabolism, the subnetworks for the production of methionine include parts of the one carbon metabolism and therefore suggest three different carbon precursors from the core metabolism for the methyl group. Furthermore, the cysteine precursor in the reference model is replaced by the cysteine precursors aspartate or 3-phosphoglycerate in our model.

In the case of tryptophan, our method proposes two lumped reactions to produce this amino acid (Figure 2.6B). The first lumped reaction uses erythrose-4-phosphate, ribose-5-phosphate and two phosphoenolpyruvates as carbon substrates. One of the phosphoenolpyruvate molecules is transformed into pyruvate, which further reacts with indole to give tryptophan. The last step is catalyzed by the enzyme tryptophanase (TRPAS2). In the second lumped reaction, one of the phosphoenolpyruvates is replaced by 3-phosphoglycerate and transformed into serine, which reacts with indole to give tryptophan. In this second case, the last step is catalyzed by the enzyme tryptophan synthase (TRPS2). The reference model only uses the route via serine, without considering the action of TRPAS2. Furthermore, the reference model uses the conversion of glutamine to glutamate as a nitrogen source, while our lumped reaction integrates ammonia via the action of the enzyme glutamine synthetase (GLNS).

Interestingly, we found that the differences in labeling patterns appearing in category C are dependent on the level of subnetwork generation. The most basic set of subnetworks, *Smin*, contains 21 reconstructed subnetworks. Only methionine has two subnetworks at this level (*S1_met* and *S2_met*), one using aspartate and the other using 3-phosphoglycerate as carbon source for the S-methyl group. All of the other BBBs are represented by exactly one subnetwork.

The next level of minimal network, *Smin+1*, generates 5 new subnetworks. The alternative subnetwork for glycine (*S3_gly*) adds 3-phosphoglycerate as carbon source for glycine biosynthesis. Also, the new subnetwork for methionine (*S3_met*) adds formate as a possible carbon source for the S-methyl group in methionine. New subnetworks for glycine, proline and methionine (*S2_gly*, *S2_pro* and *S4_met*) are also added in *Smin+1*, but even though their carbon input and output of the lumped reactions are modified, these changes do not have any consequence on the labeling pattern of the corresponding BBBs. The next level, *Smin+2*, adds three new subnetworks; The new subnetwork for tryptophan (*S2_trp*) allows 3-phosphoglycerate instead of phosphoenolpyruvate as a carbon source for tryptophan biosynthesis. The new subnetworks for methionine (*S5_met* and *S6_met*), however, do not add any new labeling patterns in the BBBs. In the last level of subnetwork expansion, *Smin+3*, the added subnetworks for methionine (*S7_met* and *S8_met*) do not contribute to new labeling patterns for methionine either.

We conclude from this analysis that expanding the size of the subnetworks can add alternative carbon sources for BBBs, as shown in the cases of methionine, glycine and tryptophan. The labeled subnetworks from the core precursors to all the BBBs are organized in a database and visualized online. The website (<http://lcsb-databases.epfl.ch/pathways/LabelingList>) is accessible for academic use upon subscription.

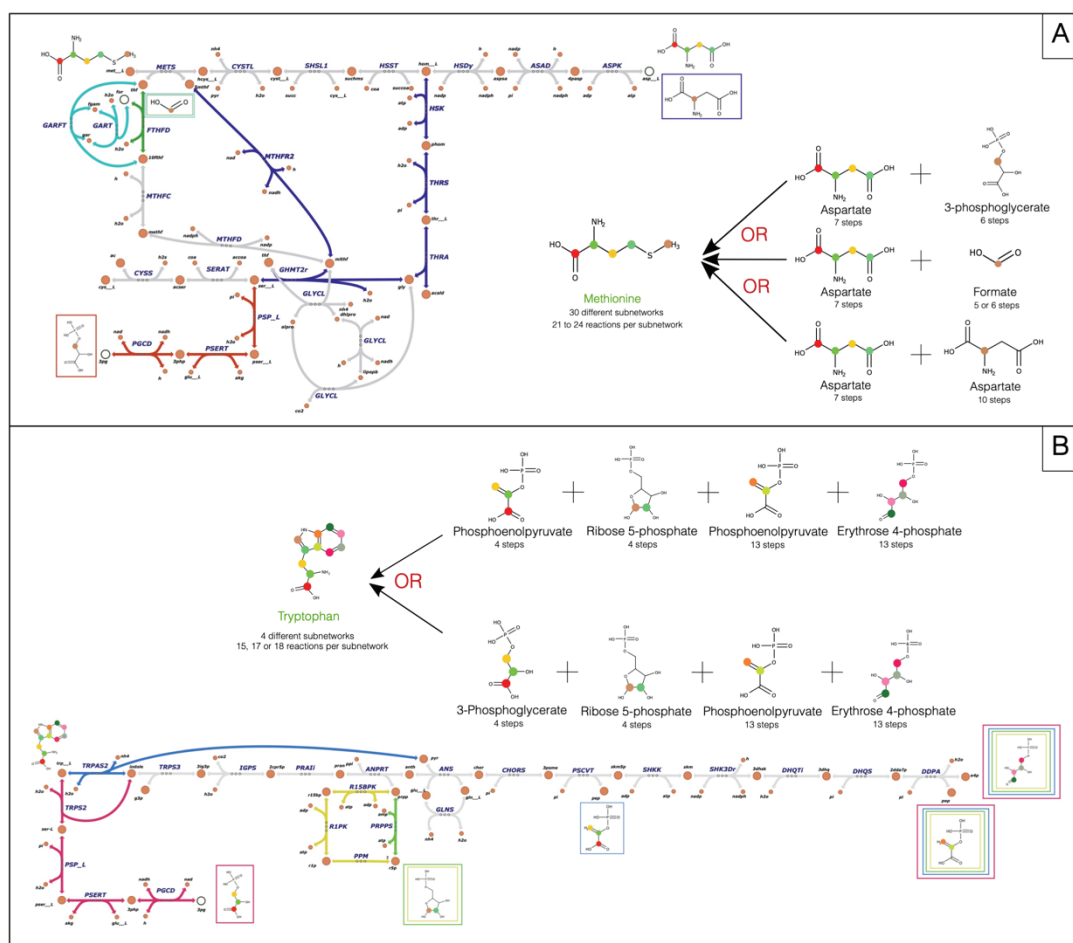


Figure 2.6: Carbon origins and labeled subnetworks for methionine (A) and tryptophan (B). Different atom positions in the amino acids are indicated with different colors. The number of reaction steps between the BBB and its labeled precursor is given. For the subnetwork visualization, different colors are used to mark the reaction that carry a label in the different subnetworks (colored edges) as well as the corresponding precursors (colored frames). (A) Blue arrows represent the subnetworks S1_met, S4_met, S5_met and S7_met, red by S2_met, and green and turquoise by S3_met, S6_met and S8_met. The turquoise subnetwork is an alternative reaction path that does is redundant with the green route in terms of labeling and final lumped stoichiometry. (B) The blue reaction maps to subnetwork S1_trp, and the pink reactions show the alternative used in S2_trp. Green and yellow arrows show alternative routes without influence on the labeling pattern. The short names of the metabolites and enzymes match the identifiers from iJO1366³⁶.

2.3.6 The redEcoli atom-level network

In the third step of the workflow, we appended the lumped reaction rules derived from the atom-mapped subnetworks to the core of redEcoli, which resulted in a full redEcoli atom-mapping model containing 92 compounds and 103 reactions. The network is divided into a core with 68 compounds and 72 reactions, and the lumped reactions derived from the carbon-mapped subnetworks with 30 lumped reactions and 24 non-core compounds. The redEcoli atom-mapping model was then used to generate carbon-mapped networks for each FDP and for each carbon position in glucose (C1 – C6) and for UTC glucose, resulting in a total of 28 carbon-mapped networks for further analysis. As an example, the carbon-mapped network for C1 without any directionality constraints on the nine BDRs had a final size of 1,159 compounds (*i.e.*, unlabeled and labeled metabolites) and 8,254 reactions (*i.e.*,

reactions carrying carbon labels, and reactions not carrying any carbon labels), out of which 6,672 were lumped reactions. This labeled example network was used to construct a hybrid model in the next step as a proof of concept.

2.3.7 Construction of an atom-level, stoichiometric hybrid model of redEcoli

The carbon-mapped redEcoli networks are used as a starting point to build a hybrid atom-level stoichiometric model of *E. coli*, which can be employed to determine the origin of carbon atoms in the Biomass Building Blocks, to predict the isotopomer distribution of ^{13}C -labeling experiments for different physiological states and to refine flux range predictions by constraining the model with ^{13}C labeling data. To create a hybrid atom-level model of redEcoli for a given labeled input substrate (*e.g.*, glucose C1 labeled), the labeled network produced in the previous step is merged into the redEcoli model (Figure 2.7). The merging process consists of the following steps: (i) All isotopomers identified with iAM.NICE are added to redEcoli. (ii) All labeled reactions are added to redEcoli. (iii) The same biochemical constraints (*i.e.*, pre-assigned directionalities) are conserved for unlabeled reactions and their corresponding labeled reactions. (iv) Uptake and secretion reactions are added for isotopomers, in case the corresponding non-labeled metabolite can be taken up and secreted in redEcoli. (v) Pooling reactions are added to combine the BBB isotopomers into a single BBB metabolite pool, which is consumed by the biomass reaction. (vi) Pooling reactions are added for all of the precursor compounds that are used to produce unlabeled BBBs, in order to ensure that unlabeled and labeled precursors flow towards BBBs that are not included in the carbon-mapped redEcoli network. This procedure ensures that the added labeling information does not interfere with the basic functionalities of the *E. coli* model. FBA is used to confirm that growth is not affected. Resulting from the merging procedure, the hybrid redEcoli model for C1 in the case of noFDP counted 1,403 metabolites and 10,235 reactions.

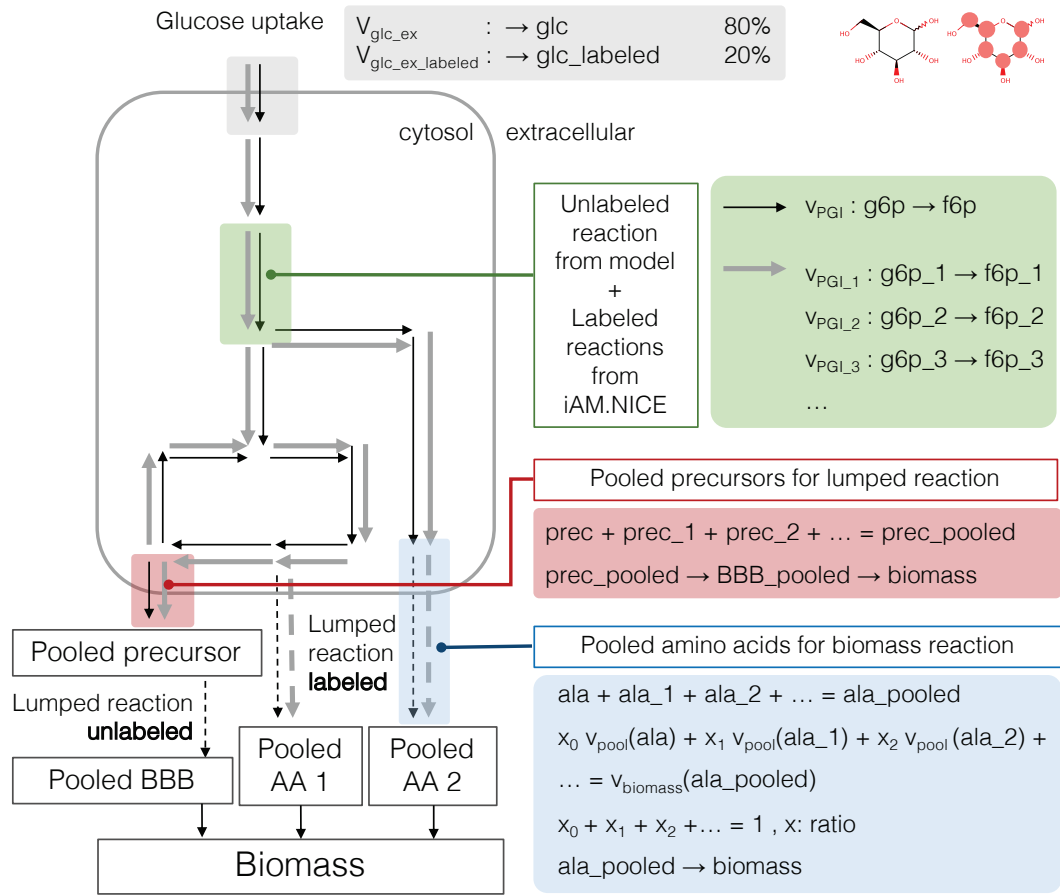


Figure 2.7: Schematic of the hybrid model of redEcoli. Thick grey arrows represent labeled reactions, thin black arrows represent unlabeled reactions, and dashed arrows represent lumped reactions. AA: amino acid, BBB: biomass building block, glc: glucose, g6p: glucose 6-phosphate, f6p: fructose 6-phosphate, PGI: phosphoglucosomerase, glc_ex: glucose exchange flux, ala: alanine, prec: precursor.

2.3.8 Refining flux ranges by incorporating experimental ^{13}C distributions

The analysis of the hybrid model is still ongoing. To benchmark our approach, we will integrate the experimental data from Leighty, R. W. & Antoniewicz⁴² into our model and compare our conclusions with their results, obtained from a state-of-the-art ^{13}C -MFA analysis. In their experiment, they grow *E. coli* on hundred percent single-labeled glucose substrates, ranging from C1 to C6. They then measured 18 amino acid fragments that arose from the fragmentation of ten different proteinogenic amino acids. For each amino acid fragment, and for each possible single-labeled glucose substrate, the percentages of labeled carbon have been determined by GC-MS. To compare with their results, we incorporated the experimentally determined ratios into the hybrid redEcoli model without directionality constraints (noFDP) at the level of the pooling flux ratios of amino acid isotopomers towards biomass amino acids. Currently, we are performing Flux Variability Analysis to determine the allowed flux ranges for each labeled core reaction. These results will enable us to

benchmark our method against the standard ^{13}C -MFA analysis, and hopefully provide new insights into the distribution of fluxes in *E. coli* in different physiological conditions.

2.3.9 Conclusions and Outlook

Here, we propose a combined framework of a biochemically correct atom tracking method with a constraint-based metabolic modeling approach for the exact modeling and analysis of ^{13}C labeling experiments. Our method overcomes the need of a top-down construction of atom-mapped core models by applying a robust, systematic reduction technique to the genome-scale model of *E. coli*. By constraining the model with thermodynamic data, we could show that pre-assumed reaction directionalities in the core model affect the interpretation of labeling experiments. We further illustrated the importance of a robust definition of lumped reactions for amino acid production. Finally, we constructed a hybrid model that allows direct integration of experimentally obtained ratios of labeled amino acids. To sum it up, the ^{13}C -FBA workflow can be used to directly constrain a metabolic model with stable-isotope labeling data, which will further help guiding experimental design in terms of tracer optimization.

While this study focuses on tracing carbon atoms, the approach can be readily used to track elements other than carbon. However, only a few experimental studies have systematically looked at the system-wide distribution and cycling of non-carbon elements such as oxygen, phosphate or sulfur in *E. coli* or other model organisms, which makes it currently difficult to benchmark predictions. However, ^{15}N -labeling has been extensively used to study nitrogen metabolism in plants⁴³, showing the interest in such studies in other organisms. Given that our approach is based on manually encoded biochemical reaction rules, and given that we validated our approach on carbon transformations, we can reasonably expect that our method can correctly trace elements other than carbon.

2.4 Tracking substrate utilization in the malaria parasite

The following Subchapter summarizes the work accomplished by master student Beatriz Lopes under the co-supervision of Professor Nuno Gonalo Pereira Mira from the Instituto Superior T cnico, Lisboa and available on the University website (https://fenix.tecnico.ulisboa.pt/downloadFile/1689244997257940/thesis_BeatrizLopes.pdf). The GEM of the malaria parasite has been developed by Dr. Anush Chiappino-Pepe⁴⁴, and her master student Thomas Gordon-Lennox performed the reduction of the GEM. The purpose of this Subchapter is illustrative, since the main work has been presented as a master thesis.

For some organisms, it is particularly difficult to measure metabolic activity experimentally. Such is the case for the parasite *Plasmodium falciparum*, a protist living inside two hosts, human and mosquitos, and the major cause for malaria⁴⁵. Studying *P. falciparum* is difficult for multiple reasons. First of all, the parasite changes its physiology drastically throughout the different stages of its life cycle. The parasite enters the human system in the form of sporozoites through the bite of an infected mosquito. From the blood stream, the parasite reaches the liver where it multiplies, before being released again into the blood stream where it infects the red blood cells (*i.e.*, erythrocytes) for reproduction. The so-called the blood stage is the symptomatic part of the infection. Another reason why the metabolism is difficult to study is its close and complex interaction with the host, which makes it difficult to measure metabolic properties (*e.g.*, consumption rate of different substrates) of the obligate intracellular parasite. These factors make it difficult to study the *P. falciparum* and to develop new drug targets, which are urgently needed given the emergence of strains that are resistant to the standard artemisinin-based treatment⁴⁶. Computational approaches are therefore key to understand the metabolism of the parasite and to identify metabolic functions that are essential for its survival. These essential functions are potential drug targets that are important for the development of new medicines⁴⁷.

In a recent substrate essentiality study, it has been shown that *P. falciparum* can compensate the lack of certain substrates if similar substrates containing the essential molecular moieties are present in the media (Figure 2.8)⁴⁴. However, to understand how the parasite metabolizes different substrates, and which molecular moieties are important for the parasite to sustain itself and reproduce, an atom-level model of the parasite was thought to provide deeper insights. On this basis, we decided to build an atom-mapping model of *P. falciparum* in the infectious blood stage of the parasite. The two objectives of this project were (i) to reconstruct an atom-level, genome-scale metabolic network that would allow us to track single atoms throughout the metabolism of *P. falciparum* and (ii) to study the substrate utilization of the parasite by tracking single carbon atoms of different substrates. To achieve this, we first constructed a reduced atom-level model of *P. falciparum*, which was further used to track single carbon atoms throughout the metabolic network. The qualitative analysis of carbon distribution revealed new insights into the carbon metabolism of *P. falciparum*, and we could show the usefulness of iAM.NICE to analyze the metabolism of non-model organisms.

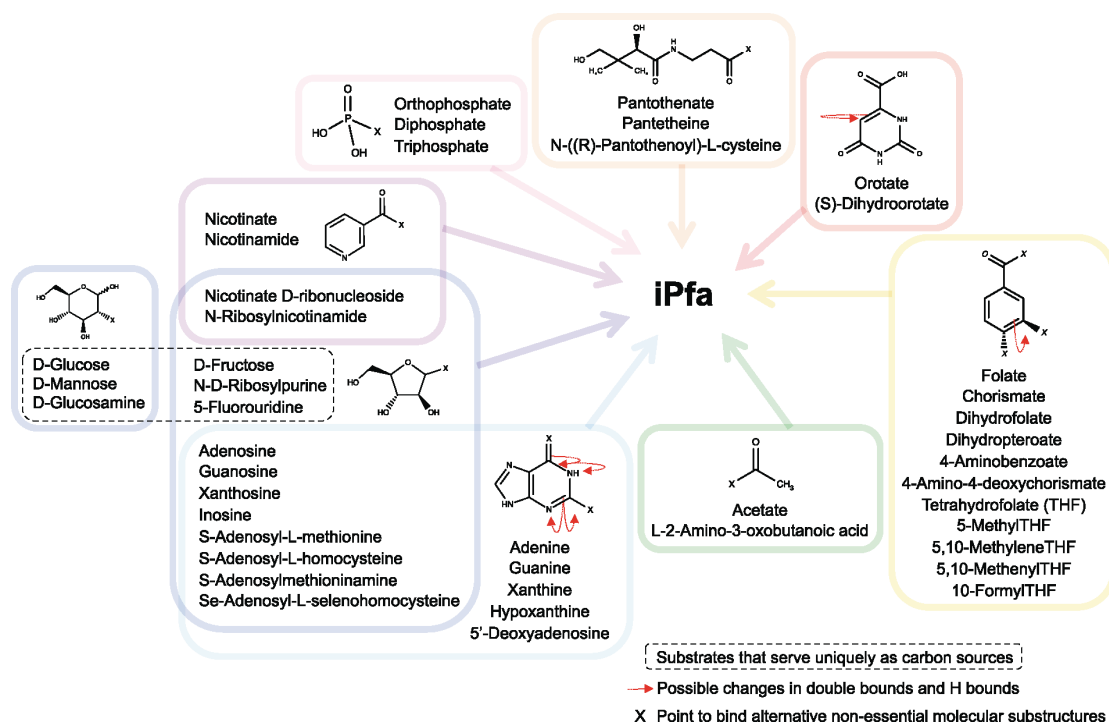


Figure 2.8: Substrate metabolites can substitute each other in the media in iPfa. Each essential substructure, or moiety, can be obtained from a range of alternative substrates. Adapted from Chiappino-Pepe et al.⁴⁴.

2.4.1 Tracing substrate atoms in a reduced model of *P. falciparum*

To study the substrate utilization of *P. falciparum* at the atomic level, we chose the genome-scale metabolic model iPfa⁴⁴, which has 1066 reactions and 1258 metabolites and consists of five distinct compartments (*i.e.*, cytosol, nucleus, mitochondrion, endoplasmic reticulum, apicoplast). In iPfa, the metabolic objective is defined as the production of 73 BBBs that are essential for the survival of the parasite. In the blood stage, the erythrocytes provide a rich medium to the parasites, meaning that all of the twenty proteinogenic amino acids and nine additional BBBs can be directly obtained from the human blood cell.

One crucial point for studying non-model organisms at the atomic level is a consistent reduction of the genome-scale model. Organizing its complexity helps to focus the atom-level studies on the most important parts of metabolism, while keeping intact the properties of the genome-scale model. iPfa was reduced using redGEM³⁷ and lumpGEM³⁸ around the sub-systems glycolysis, TCA cycle, pyruvate metabolism, pentose phosphate pathway, electron transport chain, isoprenoid metabolism, redox metabolism and hemoglobin digestion. The resulting reduced model, called rediPfa, is specific to the blood stage of the parasite. It consists of a core network of 365 metabolites and 341 reactions, plus 61 lumped reactions that produce 41 BBBs from core precursors. Out of the remaining BBBs, three are produced in the core, nine are directly taken up from the host (*e.g.*, lipoate, thiamine, choline, fatty acids), and twenty can be obtained from hemoglobin digestion (*i.e.*, amino acids).

To obtain an atom-level description of *P. falciparum*, rediPfa was annotated with BNICE.ch reaction mechanisms. 80% of the metabolic reactions in rediPfa could be reconstructed with

reaction rules, which allowed the production of all of the 44 produced BBBs in rediPfa. The remaining reactions were not added to the atom level reconstruction, since they were not interfering with the carbon labeling of the BBBs, such as electron transfers between cofactors and reactions involving compounds without defined structure.

In a first study, we simulated ^{13}C tracer experiments by labeling carbon atoms in the three different carbon sources glucose, glucosamine and fructose. Based on the experimental observation that the parasite secretes amino acids in the blood stage, we chose two different FDPs within rediPfa that allowed at least 90% of the maximal growth, one maximizing the secretion of amino acids by the cell, and one minimizing it. We further performed several *in silico* labeling experiments on the reduced model by labeling carbon atoms in the three different substrates and observing their fate in the different metabolites. Analyzing the different labeling distribution for different combinations of flux directionality profiles and labeling scenarios showed, for example, that only the carbons originating from glucose were metabolized through the pentose phosphate pathways. We also found that maximizing the secretion of amino acids resulted in a higher number of generated isotopomers, which suggests that the three carbon sources were preferentially used to produce the different BBBs, while the amino acids obtained from the degradation of hemoglobin were rather directly secreted by the cell instead of entering the central carbon metabolism. As a conclusion, we could show that directionality assumptions significantly affected the distribution of the carbon labels within the parasite.

In a second study, we traced the atoms from core precursors to corresponding BBBs. For example, we identified six different BBBs that carried the carbon labels originating from guanosine monophosphate (GMP) (Figure 2.9). In this case, we could show that the carbon atoms in the six-member ring of the nucleobase guanine conserve their positions, while the ribose moiety is transformed in the BBBs derived from tetrahydrofolate. This type of analysis allowed us to automatically identify the origins of the carbon atoms of the different BBBs in iPfa.

2.4.2 Conclusions

In this work, we could show that the proposed framework for atom-tracing in combination with network reduction can be applied to study organisms other than *E. coli*. However, there are multiple challenges associated with modeling the metabolism of eukaryotes. The existence of different cellular compartments, changing cellular physiology in time (*i.e.*, life cycle) and space (*i.e.*, different organs), sparser availability of experimental data and an increased number of knowledge gaps make *in silico* atom-tracing more complex. In this study, we did not differentiate between the different cellular compartments for atom-mapping and we focused on a single stage in the life cycle of the parasite to simplify the problem. However, future developments should consider these particularities to obtain more coherent atom-level models for eukaryotes. Finally, we believe that the standardized reduction and organization of the GEM is essential to successfully model the metabolism of eukaryotes at the level of atoms.

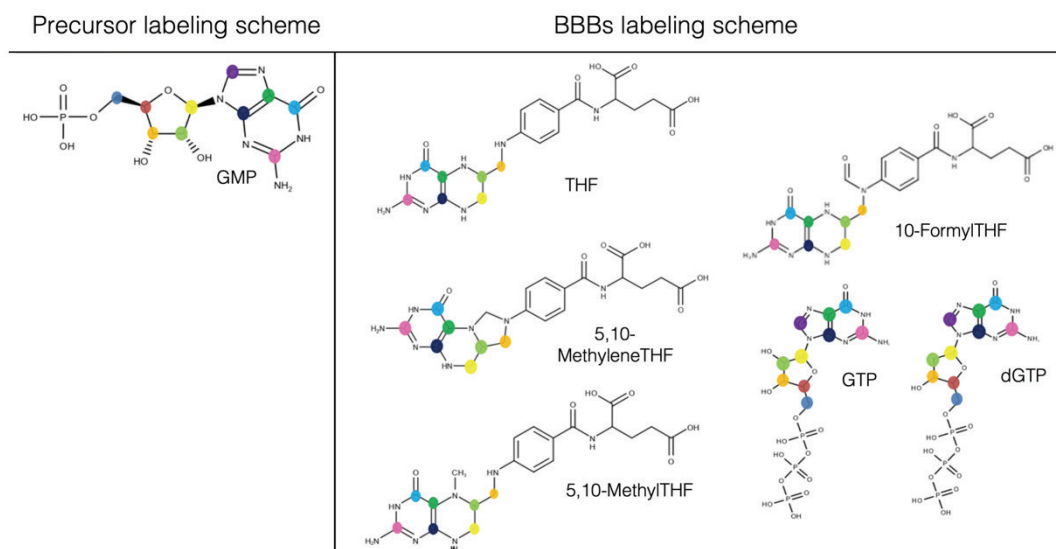


Figure 2.9: Example output of a substrate utilization study on guanosine monophosphate (GMP). Different colors designate the corresponding atomic positions in the molecular structures of the BBBs. THF: Tetrahydrofolate, GTP: Guanosine triphosphate.

2.5 Conclusion and outlook

In this chapter, we extensively explored the potential of the atom-level descriptions of enzymatic reaction rules to provide an atom-level resolution of metabolic reactions, pathways and networks. The first achievement has been to show that BNICE.ch can provide biochemically correct atom maps for reactions automatically. On the pathway level, we illustrated how walking tracing atoms backwards from secondary metabolites to core metabolites could reveal the atomic origin of the molecules. Finally, we postulate that tracing atoms in systematically reduced GEMs could provide a solid framework to analyze and model isotopic labeling experiments for a broad range of organisms. Even though ^{13}C -FBA will need further testing, development and investigation, we believe that our approach can improve the current interpretation and application of stable-isotope labeling experiments.

One outcome of the atom-mapping and -tracing studies is also educational: Acquiring a sense for metabolism at atomic resolution fueled the development of further tools and methods, as will be detailed in the following chapters. In particular, the concept of atom-mapping and -conservation within metabolic pathways has greatly helped to provide better computational solutions to the pathway search problem in big metabolic networks.

References

1. Nielsen, J. It is all about metabolic fluxes. *J. Bacteriol.* **185**, 7031–5 (2003).
2. Tang, Y. J. *et al.* Advances in analysis of microbial metabolic fluxes via ¹³C isotopic labeling. *Mass Spectrom. Rev.* **28**, 362–375 (2009).
3. Wiechert, W. ¹³C Metabolic Flux Analysis. *Metab. Eng.* **3**, 195–206 (2001).
4. Antoniewicz, M. R. ¹³C metabolic flux analysis: optimal design of isotopic labeling experiments. *Curr. Opin. Biotechnol.* **24**, 1116–1121 (2013).
5. Backman, T., Ando, D., Singh, J., Keasling, J. & García Martín, H. Constraining Genome-Scale Models to Represent the Bow Tie Structure of Metabolism for ¹³C Metabolic Flux Analysis. *Metabolites* **8**, 3 (2018).
6. McAtee, A. G., Jazmin, L. J. & Young, J. D. Application of isotope labeling experiments and ¹³C flux analysis to enable rational pathway engineering. *Curr. Opin. Biotechnol.* **36**, 50–56 (2015).
7. Gopalakrishnan, S. & Maranas, C. D. ¹³C metabolic flux analysis at a genome-scale. *Metab. Eng.* **32**, 12–22 (2015).
8. García Martín, H. *et al.* A Method to Constrain Genome-Scale Models with ¹³C Labeling Data. *PLOS Comput. Biol.* **11**, e1004363 (2015).
9. Wiechert, W., Möllney, M., Isermann, N., Wurzel, M. & de Graaf, A. A. Bidirectional reaction steps in metabolic networks: III. Explicit solution and analysis of isotopomer labeling systems. *Biotechnol. Bioeng.* **66**, 69–85 (1999).
10. Orth, J. D., Thiele, I. & Palsson, B. Ø. O. What is flux balance analysis? *Nat. Biotechnol.* **28**, 245–248 (2010).
11. Henry, C. S., Broadbelt, L. J. & Hatzimanikatis, V. Thermodynamics-Based Metabolic Flux Analysis. *Biophys. J.* **92**, 1792–1805 (2007).
12. Ataman, M. & Hatzimanikatis, V. Heading in the right direction: Thermodynamics-based network analysis and pathway engineering. *Curr. Opin. Biotechnol.* **36**, 176–182 (2015).
13. Chen, X., Alonso, A. P., Allen, D. K., Reed, J. L. & Shachar-Hill, Y. Synergy between ¹³C-metabolic flux analysis and flux balance analysis for understanding metabolic adaption to anaerobiosis in *E. coli*. *Metab. Eng.* **13**, 38–48 (2011).
14. Suthers, P. F. *et al.* Metabolic flux elucidation for large-scale models using ¹³C labeled isotopes. *Metab. Eng.* **9**, 387–405 (2007).
15. Gopalakrishnan, S. & Maranas, C. D. ¹³C metabolic flux analysis at a genome-scale. *Metab. Eng.* **1**, 12–22 (2015).
16. Hadadi, N., Hafner, J., Soh, K. C. & Hatzimanikatis, V. Reconstruction of biological pathways and metabolic networks from in silico labeled metabolites. *Biotechnol. J.* **12**, 1600464 (2017).

17. Shimizu, Y., Hattori, M., Goto, S. & Kanehisa, M. Generalized Reaction Patterns for Prediction of Unknown Enzymatic Reactions. *Genome Informatics* **20**, 149–158 (2008).
18. Chen, W. L., Chen, D. Z. & Taylor, K. T. Automatic reaction mapping and reaction center detection. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **3**, 560–593 (2013).
19. Bunke, H. On a relation between graph edit distance and maximum common subgraph. *Pattern Recognit. Lett.* **18**, 689–694 (1997).
20. Lynch, M. F. & Willett, P. The Automatic Detection of Chemical Reaction Sites. *J. Chem. Inf. Model.* **18**, 154–159 (1978).
21. Körner, R. & Apostolakis, J. Automatic Determination of Reaction Mappings and Reaction Center Information. 1. The Imaginary Transition State Energy Approach. *J. Chem. Inf. Model.* **48**, 1181–1189 (2008).
22. Kumar, A. & Maranas, C. D. CLCA: Maximum Common Molecular Substructure Queries within the MetRxn Database. *J. Chem. Inf. Model.* **54**, 3417–3438 (2014).
23. Kraut, H. *et al.* Algorithm for Reaction Classification. *J. Chem. Inf. Model.* **53**, 2884–2895 (2013).
24. Raymond, J. W. & Willett, P. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput. Aided. Mol. Des.* **16**, 521–533 (2002).
25. Arita, M. The metabolic world of Escherichia coli is not small. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 1543–1547 (2004).
26. Ravikirithi, P., Suthers, P. F. & Maranas, C. D. Construction of an E. Coli genome-scale atom mapping model for MFA calculations. *Biotechnol. Bioeng.* **108**, 1372–1382 (2011).
27. Akutsu, T. Efficient Extraction of Mapping Rules of Atoms from Enzymatic Reaction Data. *J. Comput. Biol.* **11**, 449–462 (2004).
28. Latendresse, M., Malerich, J. P., Travers, M. & Karp, P. D. Accurate Atom-Mapping Computation for Biochemical Reactions. *J. Chem. Inf. Model.* **52**, 2970–2982 (2012).
29. Heinonen, M., Lappalainen, S., Mielikäinen, T. & Rousu, J. Computing Atom Mappings for Biochemical Reactions without Subgraph Isomorphism. *J. Comput. Biol.* **18**, 43–58 (2011).
30. Crabtree, J. D., Mehta, D. P. & Kouri, T. M. An Open-Source Java Platform for Automated Reaction Mapping. *J. Chem. Inf. Model.* **50**, 1751–1756 (2010).
31. First, E. L., Gounaris, C. E. & Floudas, C. A. Stereochemically Consistent Reaction Mapping and Identification of Multiple Reaction Mechanisms through Integer Linear Optimization. *J. Chem. Inf. Model.* **52**, 84–92 (2012).
32. Fooshee, D., Andronico, A. & Baldi, P. ReactionMap: An Efficient Atom-Mapping Algorithm for Chemical Reactions. *J. Chem. Inf. Model.* **53**, 2812–2819 (2013).

33. Latendresse, M., Krummenacker, M. & Karp, P. D. Optimal metabolic route search based on atom mappings. *Bioinformatics* **btu150** (2014). doi:10.1093/bioinformatics/btu150
34. Heath, A. P., Bennett, G. N. & Kavraki, L. E. Finding metabolic pathways using atom tracking. *Bioinformatics* **26**, 1548–1555 (2010).
35. Tervo, C. J. & Reed, J. L. MapMaker and PathTracer for tracking carbon in genome-scale metabolic models. *Biotechnol. J.* **11**, 648–661 (2016).
36. Orth, J. D. *et al.* A comprehensive genome-scale reconstruction of Escherichia coli metabolism--2011. *Mol. Syst. Biol.* **7**, 535–535 (2014).
37. Ataman, M., Hernandez Gardiol, D. F., Fengos, G. & Hatzimanikatis, V. redGEM: Systematic reduction and analysis of genome-scale metabolic reconstructions for development of consistent core metabolic models. *PLoS Comput. Biol.* **13**, 1–22 (2017).
38. Ataman, M. & Hatzimanikatis, V. lumpGEM: Systematic generation of subnetworks and elementally balanced lumped reactions for the biosynthesis of target metabolites. *PLOS Comput. Biol.* **13**, e1005513 (2017).
39. Henry, C. S., Broadbelt, L. J. & Hatzimanikatis, V. Thermodynamics-based metabolic flux analysis. *Biophys. J.* **92**, 1792–1805 (2007).
40. Salvy, P. *et al.* pyTFA and matTFA: a Python package and a Matlab toolbox for Thermodynamics-based Flux Analysis. *Bioinformatics* **35**, 167–169 (2018).
41. Hatzimanikatis, V. *et al.* Exploring the diversity of complex metabolic networks. *Bioinformatics* **21**, 1603–1609 (2005).
42. Leighty, R. W. & Antoniewicz, M. R. COMPLETE-MFA: Complementary parallel labeling experiments technique for metabolic flux analysis. *Metab. Eng.* **20**, 49–55 (2013).
43. Yoneyama, T., Ito, O. & Engelaar, W. M. H. G. Uptake, metabolism and distribution of nitrogen in crop plants traced by enriched and natural ¹⁵N: Progress over the last 30 years. *Phytochem. Rev.* **2**, 121–132 (2003).
44. Chiappino-Pepe, A., Tymoshenko, S., Ataman, M., Soldati-Favre, D. & Hatzimanikatis, V. Bioenergetics-based modeling of Plasmodium falciparum metabolism reveals its essential genes, nutritional requirements, and thermodynamic bottlenecks. *PLOS Comput. Biol.* **13**, e1005397 (2017).
45. Cowman, A. F., Healer, J., Marapana, D. & Marsh, K. Malaria: Biology and Disease. *Cell* **167**, 610–624 (2016).
46. World Health Organization. *World malaria report 2015*. (2016).
47. Stanway, R. R. *et al.* Genome-Scale Identification of Essential Metabolic Processes for Targeting the Plasmodium Liver Stage. *Cell* **179**, 1112–1128.e26 (2019).

Chapter 3 Atom-conserving metabolic pathway search

During the work on Chapter 2, it became clear that the mapping and tracing of single atoms is of particular importance in metabolism, and that the information of atomic flow can help us better understand certain concepts, such as the concept of a metabolic pathway. The term “pathway” has been used in metabolic research since the first biochemical studies to describe consecutive biotransformations of metabolites. A key characteristic of a pathway is that a given property is conserved from an initial substrate to a final product. Usually, this property is a molecular substructure, but it can also be electrons or protons in specific cases such as the electron transport chain. Hence, if we know how to track the atoms of conserved substructures in a metabolic network, we can use this knowledge to retrieve metabolic pathways from biochemical networks.

Indeed, finding biotransformation pathways in biochemical networks is an important challenge in metabolic engineering. For example, one would like to produce a molecule of interest in cell chassis. In order to find all the possible pathways that lead from a precursor compound, native to the chassis, to the target molecule to be engineered, a computational pathway search is recommended to list all the possible alternative pathways that can be extracted from a biochemical database. This database may consist of known reactions (*e.g.*, KEGG), but it can also include novel, predicted reactions as created by retrobiosynthesis tools such as BNICE.ch.

The following chapter will be submitted with the title “Finding metabolic pathways in large networks through atom-conserving substrate-product pairs”. Since all of the presented work has been done by the author, no contribution statement was added to this chapter. The tool developed here, named NICEpath, is currently hosted on the code sharing platform c4science.

3.1 The quest for metabolic pathways

Extracting meaningful metabolic pathways from large metabolic networks is essential for the computational design of bioproduction pathways, for the elucidation of biosynthesis of natural products, and for the fundamental understanding of metabolism. These metabolic pathways, which describe the transformation from a source molecule over consecutive reaction steps into a target molecule, act as the roadmap guiding these various applications^{1–}

³. Traditionally, metabolic pathways were drawn by hand after directly inferring the transformations from experimental evidence. However, the advent of the omics era and the dramatic increase of computational resources has drastically changed the way we study biochemistry. Our knowledge is now collected in continuously growing databases, providing new opportunities for fundamental research and metabolic engineering, though this makes the by-hand design of pathways nearly impossible and obsolete. Additionally, these new extensive resources can be used to design non-canonical pathways that do not exist in nature. While many of these novel pathways have been historically designed by intuition using paper and pencil, it is likely that more efficient solutions will be missed. To address this challenge, computational pathway search tools have been developed to extract metabolic pathways from biochemical databases^{4,5}.

Overall, a biosynthesis pathway converts one or several metabolites into a final target metabolite containing all or most of the atoms found in the precursor compound or converts a more complex metabolite back into simpler building blocks that conserve the atoms of the original metabolite. Hence, unless stated otherwise, atom conservation between start and end point defines a metabolic pathway. If we want to find biosynthetic or biodegradation pathways, we therefore have to look for atom-conserving reaction paths connecting start and end metabolite(s). The objective of pathway search methods is to find “biologically meaningful” pathways, which are here defined as biochemical routes that fulfill the following criteria: (i) Core atoms are conserved throughout the pathway; (ii) loops are not allowed, meaning that no metabolite appears twice; and (iii) other metabolites that contribute to the main biotransformation route in a lesser degree are considered as cofactors or co-substrates. Expressed in a more general way, we aim to recover linear metabolic pathways as they are shown in textbooks, but through an automated approach.

3.1.1 Graph representation of metabolism

There are different ways to mathematically describe a metabolic network, *i.e.*, stoichiometric matrix or graph theory, and hence different approaches to analyzing biochemical networks and finding pathways⁴. However, only graph-based methods are suitable for large-scale applications due to their computational efficiency, so we will not consider other approaches here. Different methodologies have been developed in the past to (i) represent metabolic networks as mathematical graph structures, and (ii) to find pathways within the graph from a given source to a target metabolite. To bias a biochemically blind graph search algorithm towards biologically meaningful pathways, different approaches have been explored in the past, such as the exclusion of cofactors from the network, defining reactant pairs through the chemical similarity of compounds, and atom or substructure conservation throughout the pathway^{6,7}. Atom conservation in general, and carbon conservation specifically, have been shown to be a valuable criterion for finding biologically meaningful pathways^{8–12}.

There are several existing solutions to pathway discovery that employ the concept of atom conservation. Initially, the tracking of single atoms was used by Arita *et al.* to calculate network properties of the metabolism of *Escherichia coli*⁸. Later, atom tracking was used to improve the quality of pathway search tools by ensuring that one or several atoms were conserved throughout the pathway^{9,10,13,14}. Atom-tracking methods have been shown to find biologically relevant pathways, though the high quality comes with an increased computational cost. An alternative strategy has been pursued by the Kyoto Encyclopedia of Genes and Genomes (KEGG). Their reactions are annotated with chemical structure alignments, also called substrate-product pairs or reactant pairs (short RPAIRS). KEGG's pathway prediction server, named PathPred, uses the reactant pairs to create a searchable graph of biologically meaningful biotransformations¹⁵. Instead of tracking atoms individually, PathPred approximates the atom conservation by defining atom-conserving reactant pairs, which decreases the complexity of the path search problem. However, their classification system is based on a combination of manual curation and automatic annotation—a strategy that is not easily applicable to large biochemical networks, such as the ATLAS of Biochemistry with its more than one hundred thousand predicted reactions¹⁶. Large biochemical databases, especially those including hypothetical reactions, require reliable and computationally efficient algorithms to extract possible biochemical pathways.

3.1.2 Atom-conserving pathway search for large biochemical networks

Here, we address the challenge of efficiently searching and analyzing big biochemical networks. We propose a new method, named NICEpath, that biases the graph search towards atom-conserving pathways. To achieve this, we calculate weighted reactant-product pairs that reflect the atom conservation in each reaction, and we use the atom-conserving pairs to represent biochemical reaction networks as weighted graphs that are compatible with efficient search algorithms. The pathways found by NICEpath therefore fit the definition of “biologically meaningful” in the sense that they fulfill the three criteria mentioned earlier. The algorithm finds atom-conserving pathways first and returns a pathway list ranked according to atom conservation. NICEpath can be readily employed to extract and compare metabolic pathways from overall biochemical database (*e.g.*, KEGG) or from metabolic networks specific to an organism (*e.g.*, genome-scale models). The method can be further applied to efficiently search large biochemical networks, as they are generated by reaction prediction tool such as BNICE.ch.

3.2 The NICEpath method

Our approach can be divided into four steps (Figure 3.1): (i) The first step consists of acquiring an atom-level representation of each reaction. The atom maps can come from databases, atom-mapping algorithms, or, in our case, enzymatic reaction rules. (ii) In a second step, each atom-mapped reaction is decomposed into all the possible reactant-product pairs. For each pair, we calculate the Conserved Atom Ratio (CAR) from the number of conserved atoms between reactant and product and the size of the molecules in terms of number of atoms. (iii) The atom-weighted substrate-product pairs are used to construct a weighted undirected graph, where the distance between reactants and products are inversely proportional to the CAR. (iv) Once the graph of weighted substrate-product pairs is constructed, we can apply well-established graph search methods to find the shortest paths, which will naturally represent the pathways that conserve the highest number of atoms. NICEpath uses the Yen's k-shortest loop-less path¹⁷ algorithm, a standard method to find a given number of shortest paths in weighted graph, avoiding the repetition of nodes.

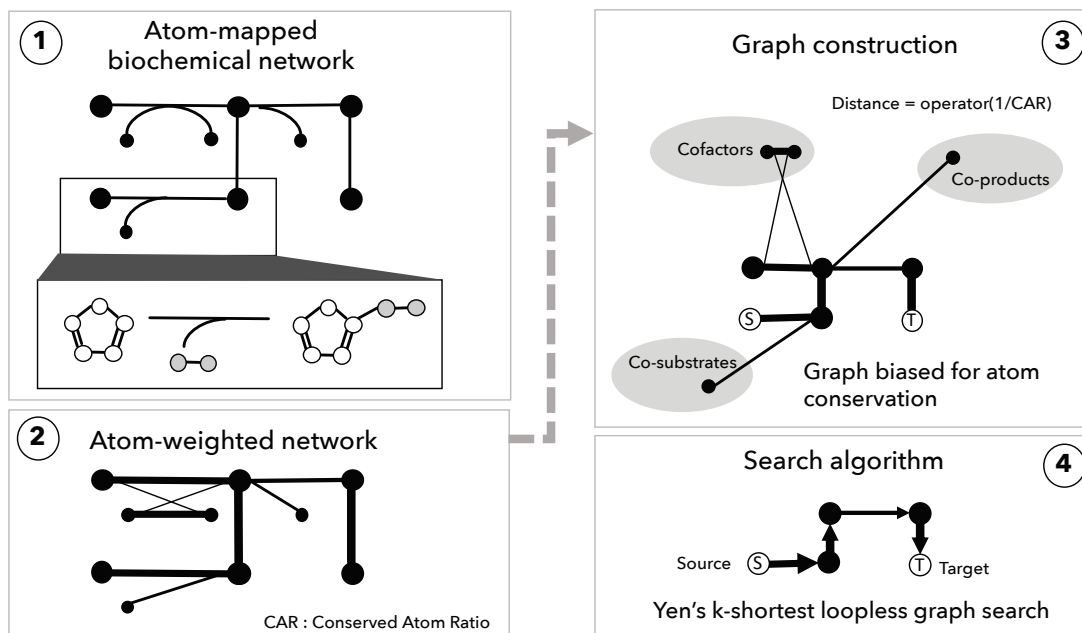


Figure 3.1: The workflow of the pathway search is divided in two parts. The first two steps (left) describe the atom weighting of the network from atom-mapped reactions. In this study, steps 1 and 2 are performed by BNICE.ch. Steps 3 and 4 (right), implemented in NICEpath, take the atom-weighted network as an input to create a searchable graph structure and finally apply a Yen's k-shortest pathway search.

3.2.1 Biochemically correct atom-mapping with BNICE.ch

Atom-mapped reactions are the prerequisite for calculating weighted reactant-product pairs. Here, we use the computational tool BNICE.ch, developed to predict hypothetical biochemical networks, to calculate biochemically correct atom mappings of enzymatic reactions. The core of BNICE.ch consists of generalized biochemical reaction rules that describe the biochemical reaction mechanisms of enzymatic reactions. The reaction rules are applied to a molecular structure to (i) reconstruct atom-mapped, known biochemical reactions; and

(ii) to predict all possible biochemical transformations that a given compound can undergo along with the product compounds generated in the process. Here, BNICE.ch calculates atom maps for metabolic reactions using the mechanistic knowledge stored in the reaction rules, as described by Hadadi *et al*¹⁸ and discussed in Chapter 2. At this step, other tools for the automatic atom mapping of reactions may also be applied to generate atom maps^{19–21}.

3.2.2 Calculation of weighted reactant-product pairs

The following steps are applied to each reaction in the network to generate atom-weighted reactant-product pairs: (i) Each reaction is split into all possible reactant-product pairs. (ii) For each pair of reactant and product, the number of common atoms (n_c) between reactant and product is calculated along with the total number of atoms in the reactant (n_r) and the total number of reactants in the product (n_p). Hydrogen atoms are omitted from the calculation. (iii) For each pair, the ratio of conserved atoms (in the following, called Conserved Atom Ratio, or CAR) is calculated with respect to the reactant (CAR_r) and with respect to the product (CAR_p).

$$CAR_r = \frac{n_c}{n_r}, \quad CAR_p = \frac{n_c}{n_p}$$

(iv) To calculate a bidirectional CAR, the mean CAR is multiplied with a correction factor that increases with the size difference between the number of common atoms and the total number of atoms in the molecule.

$$CAR = \frac{CAR_r + CAR_p}{2} \cdot (1 - |CAR_r - CAR_p|)$$

The only exception to this approach is made for reactions involving the cofactor Coenzyme A (CoA). In a molecule, CoA is treated as a single atom when it occurs in both the reactant and in the product, mainly because the high number of conserved atoms between the comparably big CoA leads to high CARs, thus masking the biochemically more interesting connections between the smaller metabolites that are attached to and detached from CoA during metabolic transformations. The final CAR value is used to weight reactant-product pairs in the network.

3.2.3 Assigning mechanisms to biochemical reactions from the KEGG reference network

We used KEGG as a reference database for enzymatic reactions, from which we extracted all reactions that have an associated mechanism in BNICE.ch. If a given reaction from KEGG could be reconstructed with BNICE.ch, it was assigned a reaction mechanism that allowed us to retrieve the number of conserved atoms between each reactant-product pair. The set of KEGG reactions with assigned reaction mechanisms and pre-calculated CAR values was used for further validation and as an example network for network analysis and pathway search. The set of BNICE.ch curated KEGG reactions will be available from the NICEPath repository.

3.2.4 Graph representation of biochemical networks

For a given reaction network, NICEpath loads all the reactant-product pairs to generate a weighted, undirected graph, where metabolites are nodes connected by edges, representing the reactant-product relationship. Edges are assigned a weight that defines the relation between two connected nodes. To use state-of-the-art shortest-path graph search algorithms, highly atom-conserving reactants should be close to each other, and pairs that only share a few atoms should be further away. Hence, we convert the CAR into a distance:

$$\text{default transformation: } \text{distance} = \frac{1}{CAR}$$

NICEpath accepts two alternative ways to calculate the distance, which can be used to modulate the influence of the atom conservation on the weight of the reactant-product pair.

$$\text{square root transformation: } \text{distance} = \sqrt[2]{\frac{1}{CAR}}$$

$$\text{exponential transformation: } \text{distance} = \frac{e^{1/CAR}}{e}$$

The type of transformation can be changed to square root or exponential depending on the nature of the pathway search problem, *i.e.*, the structures of source and target molecules as well as the estimated number of biotransformation used to convert one into the other. The distance measure is used to reconstruct a directed graph whose edge weights represent the atomic distance between reactants and products. For longer pathways, we recommend using the exponential transformation because it increases the penalty for pairs with low CARs, which makes the search more conservative in terms of atoms. For this study, we grouped duplicate KEGG compounds into one node. Duplicates were identified based on

the first fourteen letters of the InChIKey, which means that different stereoisomers of the same molecular structure were merged into one node.

3.2.5 Finding metabolic pathways with graph search

NICEpath applies a Yen's k-shortest loop-less path search to extract the shortest pathways from the weighted network of reactant-product pairs¹⁷ using the python package NetworkX. As inputs, the pathway search algorithm takes a weighted graph, a source compound, a target compound, and the maximum number of shortest paths (k) to be found. As soon as this number k is reached, the algorithm stops and returns all the k-shortest paths in terms of summed edge weights.

The run time of NICEpath depends on the structure of the network, the distance between the source and target compound in the graph, the number of pathways to be found, and the maximum pathway length allowed. As an example, to find 10,000 pathways of maximum length 100, the algorithm runs for about 15 minutes on a standard desktop computer using a single core. If there are several source compounds given as input, NICEpath runs path searches in parallel for different source compounds using all available cores.

3.2.6 Network analysis

NICEpath first calculates standard network statistics, such as the number of nodes and edges, and then extracts an undirected, unweighted network from the original network by only considering edges with a CAR higher than a given threshold. For this new network, the number of components, or disjoint graphs, is extracted, and the biggest component is further analyzed regarding its size relative to the previous network as well as its diameter. Since searching for pathways between two compounds belonging to different disconnected graphs will not yield any good pathways, NICEpath will warn the user in this case.

3.2.7 Software

The NICEpath code can be executed with any python version up to 3.7. The NetworkX python library (<https://networkx.github.io/>) was used to implement and search the reaction graph. An extensive list of libraries used will be available in the specification file in the repository.

3.3 Results and discussion

To demonstrate the utility and biological importance of our methods, this section starts with a validation of the biochemical relevance of the CAR metric. It is followed by a short graph-theoretical analysis of the KEGG reaction network and concluded by two practical examples of pathway searches within the KEGG network.

3.3.1 The CAR captures the main biotransformations

To validate the biochemical relevance of weighted substrate-product pairs, we compared them to the KEGG RPAIR database consisting of manually curated, atom-mapped, substructure-conserving reactant-product pairs, called RPAIRs²². KEGG differentiates between five types of RPAIRS: “main”, “cofac”, “trans”, “ligase”, and “leave”. The four latter ones describe cofactor pairs, small groups transferred by transferases, nucleotide triphosphate consumption by ligases, and the addition or removal of small inorganic compounds by lyases and hydrolases, respectively. The first type, “main”, describes the main biotransformation in a given reactions. To take the alcohol dehydrogenase reaction as an example, the main pair would be the transformation of the primary alcohol to the aldehyde, and the conversion of the cofactor NAD⁺ to NADH would be of type “cofac” (Figure 3.2). “Main” pairs that are used to draw the KEGG metabolic pathway maps. Therefore, a method that accurately predicts KEGG RPAIRS of type “main” can be used to reconstruct biologically relevant metabolic pathways. It should be noted at this point that KEGG discontinued the manual definition and curation of RPAIRS in 2016.

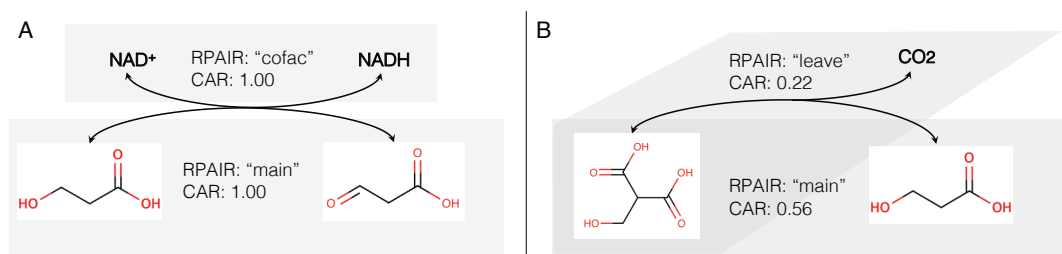


Figure 3.2: Example of relation between KEGG RPAIRs and the CAR value in a biochemical reaction. (A) Alcohol dehydrogenase, (B) decarboxylation reaction.

We validated the NICEpath method by predicting KEGG RPAIRS of type “main” using the concept of the Conserved Atom Ratio. We used BNICE.ch to calculate CAR values for a test set of 6,546 KEGG reactions for which the exact reaction mechanism is known, and which are, therefore, reconstructed by BNICE.ch. From these 6,546 reactions, we determined 10,747 substrate-product pairs with a non-zero CAR, meaning that at least one non-hydrogen atom is conserved between the substrate and the product (Appendix Table A2). Out of these 10,747 pairs, 5,148 were found to be KEGG RPAIRS of type “main”. Since RPAIRs are defined based on the conservation of structural moieties within a reaction, we hypothesized that the higher the CAR value, the more atoms conserved between a substrate and a product, and hence the higher the probability that the pair would be a KEGG RPAIR of type

“main”. We should therefore be able to predict the membership of a pair to the set of “main” KEGG RPAIRS by using a given CAR threshold as a classifier.

To test our hypothesis that the CAR is a good predictor for a reactant-product pair to be of KEGG RPAIR type “main”, we performed a Receiver-Operator Characteristic (ROC) analysis (Figure 3.3). The reference for true pairs were the 5,148 “main” RPAIRs (true positives), and the remaining 5,599 pairs were used as true negatives. For 100 CAR cutoff values between zero and one we calculated the number of good predictions (*i.e.*, number of pairs with a CAR above the cutoff and of type “main”, or true positives) and bad predictions (*i.e.*, number of pairs with a CAR above the cutoff and not of type “main”, or false positives). By drawing true positives versus false positives, we found an Area Under Curve (AUC) of 0.88. An AUC between 0.8 and 0.9 is generally considered an “excellent discrimination”²³. We further show the tradeoff between sensitivity and specificity, as well as the Youden’s index (*i.e.*, sensitivity + specificity - 1) to characterize this tradeoff²⁴ and to determine an optimal CAR cutoff. We found that the Youden’s index is maximal at a CAR equal to 0.34, which suggests that this is the optimal CAR cutoff to tell whether a given substrate-product pair conserves enough atoms to be considered a “main” pair. This analysis shows that we can reliably use the CAR to predict KEGG RPAIRS of type “main”. The network of weighted KEGG reactant pairs for 6,546 KEGG reactions is included in the NICEpath program and used as a reaction database in the default search.

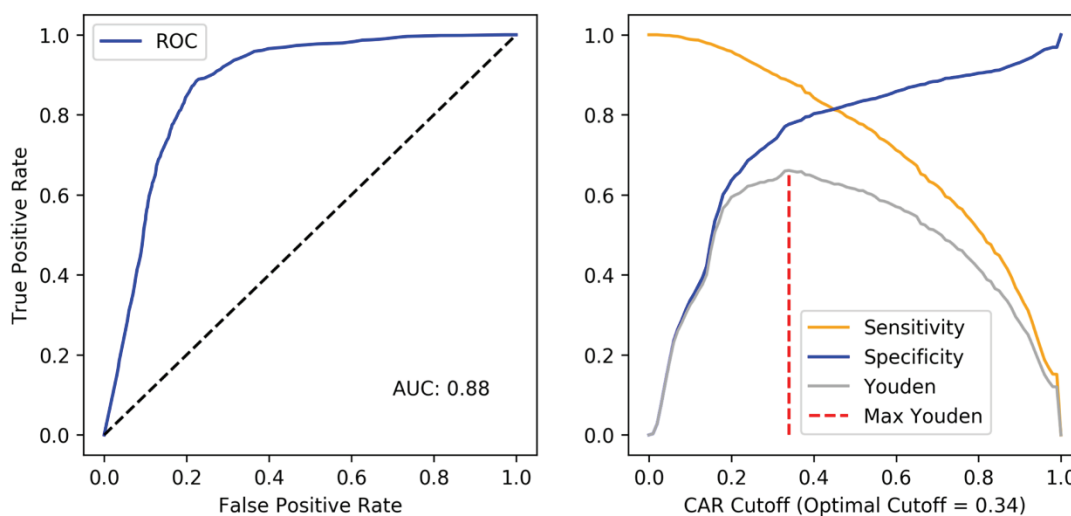


Figure 3.3: (Left) The ROC curve shows the prediction of KEGG RPAIRS of type “main” by CAR score from BNICE.ch. (Right) The trade-off between specificity (blue) and sensitivity (red). The Youden’s index (yellow) reaches its maximum (0.66) at a CAR value of 0.34.

3.3.2 Graph-theoretical analysis of metabolic networks

Characterizing biochemical networks from a graph-theoretical point of view can be used to evaluate the quality and connectivity of the represented network, and also bring new insights into the overall organization of metabolism. Furthermore, knowing the graph-theoretical properties of a biochemical network can be crucial for anticipating potential problems in the pathway search. NICEpath provides basic network statistics that allow us to

assess the quality of the data. Here, the weighted graph of the KEGG network used for validation initially contained 5,578 compounds, or nodes, and 20,911 directed edges representing reactant-product pairs.

Certain graph properties are not defined for weighted directed graphs, such as the number of components or the network diameter. For calculating these properties, a simple, non-directed graph was generated by removing reactant-product pairs with a CAR lower than the previously calculated threshold of 0.34 and by removing the weights on the remaining reactant-product pairs. The unweighted graph contains 5,518 nodes and 5,541 edges, which are distributed over 813 smaller disjoint graphs, or so-called components. The biggest component contains 2,663 nodes (48%) and 3,422 edges (62%), and it has a network diameter of 40. In other words, the longest shortest pathway connecting two compounds counts 40 biotransformation steps in the main component of the KEGG network. This means that the KEGG network is dominated by a one big component, or subnetwork, that includes half of the metabolites in KEGG and represents the core metabolism plus connected secondary metabolism. The remaining metabolites are organized in small, disconnected subnetworks, which we hypothesize to be mostly secondary metabolites without defined biosynthesis pathways.

3.3.3 Finding biologically relevant pathways with NICEpath

To illustrate the output of NICEpath, we discuss two example pathway searches. In the first example, we tried to biochemically connect tyrosine to caffeate, and we allowed a maximum number of ten pathways to be found, and only one reaction alternative was returned in case several reactions could do the same biotransformation. The pathway search resulted in ten pathways with lengths ranging from two to six consecutive reaction steps (Table 3.1). The quality of the pathway can be estimated from the pathway score and the average CAR. The pathway score sums the distances for each reactant-product pair in the pathway. The score reflects both the length of the pathway as well as the quality of atom conservation within the pathway, and it is eventually used by NICEpath to rank the paths. The average CAR estimates the quality of the pathway by averaging the atom conservation over each reaction step.

Out of these ten best pathways, the pathways ranked first, second, and fifth were chosen for visual inspection (Figure 3.4). The first pathway had a very low score of 2.24 combined with a high average CAR (0.89) and a length of two, which indicates that the pathway is of good quality because it can be synthesized in a small number of steps with high atom conservation. Indeed, KEGG proposes this pathway in the pathway map for phenylpropanoid biosynthesis, meaning that it is biologically relevant. The second pathway, although longer, has a similarly high average CAR of 0.93, a length of four steps, and it can also be found in KEGG. To contrast these two good pathway examples with a poor example, the pathway ranked fifth shows a slightly lower average CAR of 0.81, which is due to the attachment and subsequent detachment of a one-carbon unit. Out of five reaction steps, the last step is redundant with the first pathway, while the first four steps describe a detour from tyrosine

to coumarate (C00811). This last, suboptimal pathway cannot be found in the KEGG map for phenylpropanoid biosynthesis.

Table 3.1: Output of example pathway search from tyrosine (C00082) to caffeate (C01197). KEGG identifiers are used to specify compounds and reactions.

Index	Pathway length	Intermediates	Reaction IDs	Pathway score	Average CAR
1	2	C00082->C00811->C01197	R00737->R07826	2.24	0.89
2	4	C00082->C00811->C00223->C00323->C01197	R00737->R01616->R07436->R01943	4.32	0.93
3	4	C00082->C01179->C03672->C00811->C01197	R00729->R03336->R08766->R07826	4.33	0.93
4	4	C00082->C00079->C00423->C00811->C01197	R07211->R00697->R02253->R07826	4.52	0.89
5	5	C00082->C00826->C00079->C00423->C00811->C01197	R00732->R00691->R00697->R02253->R07826	6.27	0.81
6	6	C00082->C01179->C03672->C00811->C00223->C00323->C01197	R00729->R03336->R08766->R01616->R07436->R01943	6.41	0.94
7	6	C00082->C00811->C00223->C00323->C00406->C01494->C01197	R00737->R01616->R07436->R01942->R02194->R03366	6.44	0.93
8	6	C00082->C00079->C00423->C00540->C00223->C00323->C01197	R07211->R00697->R02255->R08815->R07436->R01943	6.50	0.93
9	6	C00082->C00811->C00423->C00540->C00223->C00323->C01197	R00737->R02253->R02255->R08815->R07436->R01943	6.50	0.93
10	6	C00082->C00079->C00423->C00811->C00223->C00323->C01197	R07211->R00697->R02253->R01616->R07436->R01943	6.60	0.91

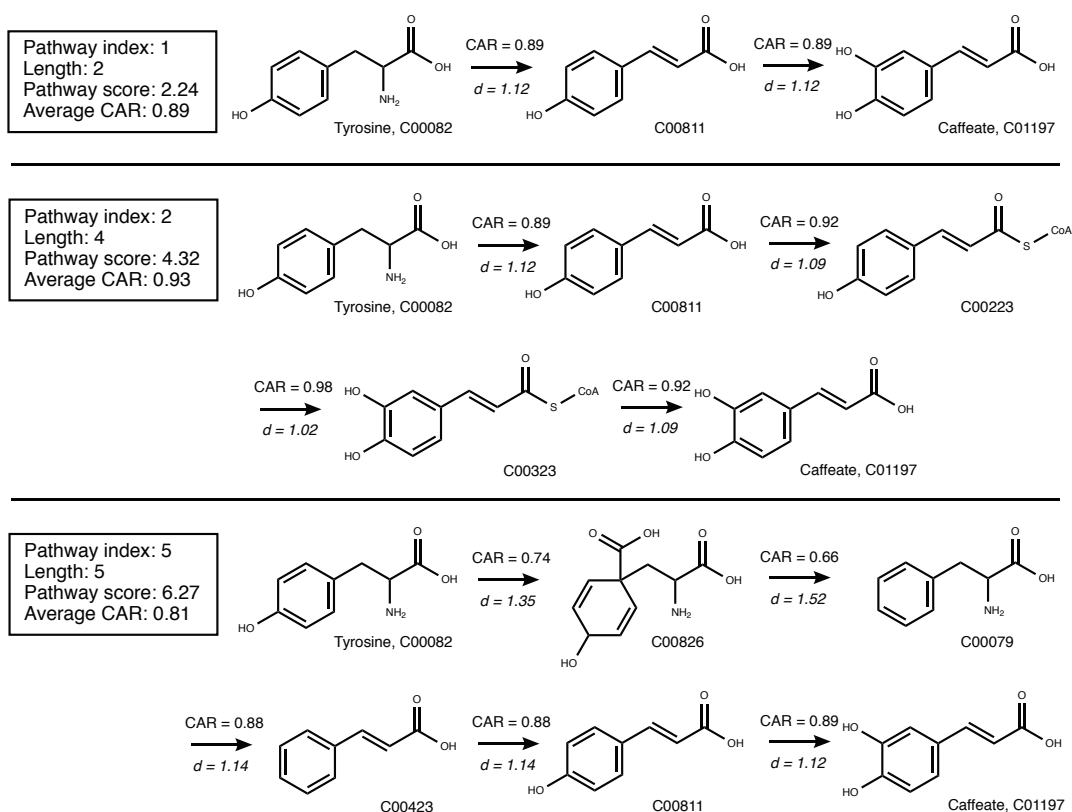


Figure 3.4: The pathways from Table 3.1 with index numbers 1, 2 and 5 connecting tyrosine and caffeate are visualized in detail for comparison. For each biotransformation, the CAR value as well as the default distance (d) are indicated.

In a second example search, we tried to find pathways connecting the compounds tyrosine and syringin. The number of pathways to be found was restricted to five, and we used three different transformations to calculate the distance between reactant-product pairs: The default transformation $1/\text{CAR}$, the square root transformation, and the exponential transformation. Using the default option, NICEpath first listed three short pathways with a low average CAR (~ 0.5), followed by two longer pathways with high average CAR (~ 0.8) (Table 3.2). The square root option yielded only short pathways with a low average CAR, while the

exponential option only resulted in longer pathways of high average CAR. Interestingly, all the long pathways with high CAR were identified as known metabolic pathways in KEGG, indicating that the exponential transformation operator is helpful to reliably search for longer pathways.

Two pathways were chosen to understand in detail the influence of the type of transformation used for calculating the distances between reactants and products: one was short with a low CAR (A) and one was long with a high CAR (D*) (Figure 3.5). Pathway A connected tyrosine to syringin in four reaction steps, with a relatively low average CAR of 0.51. As already indicated by the low CAR, the pathway turned out to be a shortcut through glucose, with no atoms conserved between tyrosine and syringin. The pathway was ranked first in the default and the square root transformation types, but, interestingly, ranked 1114th in the exponential case. The exponential transformation increases the penalty of atom loss in biotransformation, which leads to a higher pathway score assigned to the shortcut pathway. Pathway D* connected tyrosine to syringin in eight reaction steps, with a high average CAR of 0.86. It was ranked first using an exponential transformation, ranked fourth using the default distance calculation, and ranked 43rd for the square root case. This second pathway kept the molecular core structure of tyrosine and modified it to produce syringin, conserving a maximum number of atoms. Remarkably, this pathway is part of the KEGG pathway map for phenylpropanoid biosynthesis, and it can therefore be called a confirmed, biologically meaningful pathway.

These two examples of pathway search problems illustrate the capacity of NICEpath to efficiently extract biologically relevant pathways from large biochemical networks. The algorithm robustly handled searches for long pathways of eight and more biotransformation steps, as they are usually present in secondary metabolism.

Table 3.2: Output of pathway search from tyrosine (C00082) to syringin (C01197). Pathways are mapped across the different distance transformations with letters indicated in the column “Mapping”. Pathways marked with an asterisk (*) correspond to known metabolic pathways that fulfil the criteria for biologically relevant pathways.

Distance	Index	Pathway length	Intermediates	Mapping	Pathway score	Average CAR
$\frac{1}{CAR}$	1	4	C00082->C00811->C16827->C00031->C01533	A	9.06	0.51
	2	4	C00082->C00811->C04415->C00029->C01533	B	9.86	0.48
	3	4	C00082->C00811->C16827->C00029->C01533	C	9.86	0.48
	4	8	C00082->C00811->C01197->C01494->C05619->C00482->C05610->C02325->C01533	D*	9.88	0.86
	5	8	C00082->C00811->C01197->C01494->C02666->C12204->C05610->C02325->C01533	E*	9.90	0.85
$2\sqrt{\frac{1}{CAR}}$	1	4	C00082->C00811->C16827->C00031->C01533	A	5.93	0.51
	2	4	C00082->C00811->C04415->C00029->C01533	B	6.18	0.48
	3	4	C00082->C00811->C16827->C00029->C01533	C	6.18	0.48
	4	5	C00082->C00079->C00423->C04164->C00031->C01533	F	7.00	0.58
	5	5	C00082->C00811->C00423->C04164->C00031->C01533	G	7.00	0.58
$\frac{1}{e^{1/CAR}}$	1	8	C00082->C00811->C01197->C01494->C05619->C00482->C05610->C02325->C01533	D*	11.09	0.86
	2	8	C00082->C00811->C01197->C01494->C02666->C12204->C12205->C02325->C01533	H*	11.13	0.85
	3	8	C00082->C00811->C01197->C01494->C02666->C00590->C12205->C02325->C01533	I*	11.13	0.85
	4	8	C00082->C00811->C01197->C01494->C02666->C12204->C05610->C02325->C01533	E*	11.13	0.85
	5	9	C00082->C00811->C00223->C00323->C00406->C12203->C00411->C05610->C02325->C01533	J*	11.79	0.90

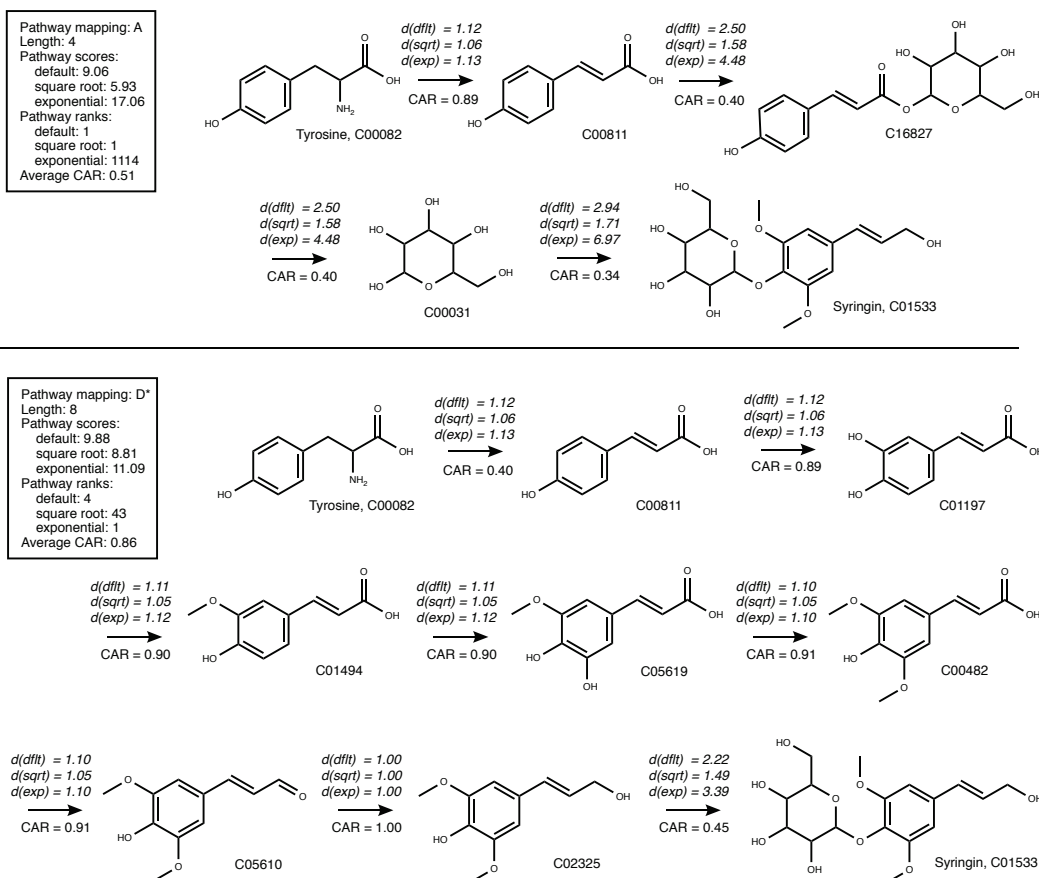


Figure 3.5: Comparison of two pathways (A and D*) from the pathway search connecting tyrosine to syringin. For each biotransformation, the CAR value as well as the default distances for each transformation are indicated. d(dflt): default distance, d(sqrt): square root transformation, d(exp): exponential transformation.

3.3.4 Limitations and future challenges

However, there are cases in which NICEpath will not find satisfactory solutions. Possible reasons for suboptimal results are (i) the network does not contain the necessary reactions to connect the starting compound to the target compound, and (ii) the source and the target compound do not initially have a lot of atoms in common. The first issue can be solved by adding the missing reactant-product pairs to the network. Missing steps can be hypothesized manually or predicted using reaction prediction tools such as BNICE.ch. The second issue is more complex, since it depends on the molecular structure of the source and target compound, as well as on the real number of biochemical transformations needed to transform one into the other. Possible solutions to improve the output include breaking down the search into several sub-searches by identifying intermediates and increasing the penalty on atom loss by using an exponential transformation of $1/\text{CAR}$ to weight the reactant-product edges in the graph.

While our algorithm successfully circumvents the recurrent problem of shortcuts through small hub metabolites, it does not satisfactorily avoid shortcuts through big hub metabolites such as Coenzyme A (CoA). In fact, reactant pairs involving CoA structures on both sides

have a lot of atoms in common, and hence a high CAR value. For this reason, NICEpath excludes CoA by default from the reactant pair network.

3.4 Conclusion

We introduce a new pathway search method based on weighted reactant-product pairs. To our best knowledge, this is the first to use automatically generated atom-weighted reactant-product pairs in combination with a k-shortest graph search approach. We benchmarked our method for reactant-pair weighting against the KEGG RPAIR database, and we discussed the advantages of NICEpath on practical examples. The strong point of NICEpath is that it is suitable for big biochemical networks, spanning more than hundreds of thousands biochemical reactions, such as hypothetical reaction networks generated by retrobiosynthesis tools and predictive biochemistry¹⁶. We estimate that the future development of reaction prediction tools, based on biochemical reaction rules or machine learning methods, will yield big hypothetical reaction networks that require optimized search tools to efficiently extract biochemical pathways.

Finally, the herein proposed framework will lay the foundation for further developments. Other types of weights, such as kinetic and thermodynamic considerations, can be integrated into the weighting of substrate-product pairs to steer the pathway search towards biochemically feasible pathways, and a set of user-defined parameters will make it easy to fine-tune the pathway search. We plan to make the NICEpath tool available on GitHub, including a collection of 5,434 known metabolic reactions with pre-calculated atom-weighted reactant pairs. The graph representation method and the pathway search developed in this Chapter will find practical application in Chapter 4 and 5 of this thesis.

References

1. Lin, G.-M., Warden-Rothman, R. & Voigt, C. A. Retrosynthetic design of metabolic pathways to chemicals not found in nature. *Curr. Opin. Syst. Biol.* (2019). doi:10.1016/J.COISB.2019.04.004
2. Nielsen, J. & Keasling, J. D. Engineering Cellular Metabolism. *Cell* **164**, 1185–1197 (2016).
3. Cravens, A., Payne, J. & Smolke, C. D. Synthetic biology strategies for microbial biosynthesis of plant natural products. *Nat. Commun.* **10**, 2142 (2019).
4. Wang, L., Dash, S., Ng, C. Y. & Maranas, C. D. A review of computational tools for design and reconstruction of metabolic pathways. *Synth. Syst. Biotechnol.* **2**, 243–252 (2017).
5. Hadadi, N. & Hatzimanikatis, V. Design of computational retrobiosynthesis tools for the design of de novo synthetic pathways. *Curr. Opin. Chem. Biol.* **28**, 99–104 (2015).
6. Sankar, A., Ranu, S. & Raman, K. Predicting novel metabolic pathways through subgraph mining. *Bioinformatics* **33**, 3955–3963 (2017).
7. Pertusi, D. A., Stine, A. E., Broadbelt, L. J. & Tyo, K. E. J. Efficient searching and annotation of metabolic networks using chemical similarity. *Bioinformatics* **31**, 1016–24 (2015).
8. Arita, M. The metabolic world of Escherichia coli is not small. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 1543–1547 (2004).
9. Heath, A. P., Bennett, G. N. & Kavraki, L. E. Finding metabolic pathways using atom tracking. *Bioinformatics* **26**, 1548–1555 (2010).
10. Huang, Y., Zhong, C., Lin, H. X. & Wang, J. A Method for Finding Metabolic Pathways Using Atomic Group Tracking. *PLoS One* **12**, e0168725 (2017).
11. Tervo, C. J. & Reed, J. L. MapMaker and PathTracer for tracking carbon in genome-scale metabolic models. *Biotechnol. J.* **11**, 648–661 (2016).
12. Kumar, A., Wang, L., Ng, C. Y. & Maranas, C. D. Pathway design using de novo steps through uncharted biochemical spaces. *Nat. Commun.* **9**, 184 (2018).
13. Latendresse, M., Krummenacker, M. & Karp, P. D. Optimal metabolic route search based on atom mappings. *Bioinformatics* btu150 (2014). doi:10.1093/bioinformatics/btu150
14. Pey, J., Planes, F. J. & Beasley, J. E. Refining Carbon Flux Paths using atomic trace data. *Bioinformatics* btt653 (2013). doi:10.1093/bioinformatics/btt653
15. Moriya, Y. *et al.* PathPred: an enzyme-catalyzed metabolic pathway prediction server. *Nucleic Acids Res.* **38**, W138–43 (2010).

16. Hadadi, N., Hafner, J., Shajkofci, A., Zisaki, A. & Hatzimanikatis, V. ATLAS of Biochemistry: A Repository of All Possible Biochemical Reactions for Synthetic Biology and Metabolic Engineering Studies. *ACS Synth. Biol.* (2016). doi:10.1021/acssynbio.6b00054
17. Yen, J. Y. Finding the K Shortest Loopless Paths in a Network. *Manage. Sci.* **17**, 712–716 (1971).
18. Hadadi, N., Hafner, J., Soh, K. C. & Hatzimanikatis, V. Reconstruction of biological pathways and metabolic networks from in silico labeled metabolites. *Biotechnol. J.* **12**, 1600464 (2017).
19. Chen, W. L., Chen, D. Z. & Taylor, K. T. Automatic reaction mapping and reaction center detection. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **3**, 560–593 (2013).
20. Latendresse, M., Malerich, J. P., Travers, M. & Karp, P. D. Accurate Atom-Mapping Computation for Biochemical Reactions. *J. Chem. Inf. Model.* **52**, 2970–2982 (2012).
21. Fooshee, D., Andronico, A. & Baldi, P. ReactionMap: An Efficient Atom-Mapping Algorithm for Chemical Reactions. *J. Chem. Inf. Model.* **53**, 2812–2819 (2013).
22. Shimizu, Y., Hattori, M., Goto, S. & Kanehisa, M. Generalized Reaction Patterns for Prediction of Unknown Enzymatic Reactions. *Genome Informatics* **20**, 149–158 (2008).
23. David W. Hosmer, Jr., S. L. *Applied Logistic Regression*. (Chapter 5, John Wiley and Sons, New York, NY, 2000).
24. Youden, W. J. Index for rating diagnostic tests. *Cancer* **3**, 32–35 (1950).

Chapter 4 ATLASx - Databases for predictive biochemistry

The ATLAS of Biochemistry is a repository of known and predicted biochemical reactions for synthetic biology and metabolic engineering applications¹. Predictive biochemistry is not only key to design biosynthesis pathways towards new bioengineering targets such as pharmaceuticals and commodity chemicals, but also to find and fill the knowledge gaps in our current understanding of metabolism. This chapter starts with an introduction on the knowledge gaps in biochemistry, followed by short review on recent advances in the field of reaction prediction (Subchapter 3.1). 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 bioATLAS and chemATLAS, as well as the future extension novATLAS (Subchapter 3.4).

4.1 “Dark matter” in metabolism

Metabolic “dark matter” designates biochemical processes that are difficult to measure and barely understood, *i.e.* underground metabolism. These underground processes are the result of the unknown or promiscuous activity of enzymes^{2,3}, and they manifest themselves in unexpected resistance to gene knock-outs, novel natural products and chemical derivatives of damage-prone metabolites⁴. These unknowns limit our general understanding of metabolism, and they further hamper the advancement of metabolic engineering applications towards the creation of sustainable cell factories. While the omics era has provided us with a huge amount of genomic, transcriptomic, proteomic and metabolomic data, linking these data to metabolic functions is lagging behind^{5,6}. For example, we know 25 % percent of proteins in *E. coli*, one of the best studied model organism, do not have a function assigned⁷, and almost 10,000 metabolites are orphan in KEGG⁸. On top of that, it is not possible to accurately assess the number of unknown promiscuous side-reactions of enzymes or yet uncharacterized compounds. Hence, inferring the physiology of a cell from omics data remains an open challenge in the modeling of metabolism, which means that our knowledge of metabolic processes remains incomplete. Biochemical assays are the ultimate solution to identify new enzymatic functions and to detect novel natural products. However, these experiments can only focus on a single process at a time, and they are long and expensive. Consequently, alternative ways are needed to systematically explore the metabolic dark matter arising from the elasticity of enzymatic catalysis in an unbiased and global approach, which can be provided by computational approaches.

4.1.1 Cheminformatic approaches

The past decades have shown increasing interest in computational solutions 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 goal is to find biosynthetic routes that produce a given target compound in a host organism^{9–13}. This problem is solved by biochemically “walking back”, reaction step by reaction step, from the target to known precursor compounds that are native to the host organism. This procedure is called *retrobiosynthesis* and implemented in a range of tools such as BNICE.ch^{14,15}, RetroPath2.0^{16,17}, NovoStoic¹⁸, ReactPRED¹⁹. 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 molecular atom-rearrangement performed by the enzyme. By applying the rule on a substrate that is non-native to the represented enzyme, the rule can predict if the substrate can be recognized by the enzyme, if the biotransformation can occur, and what will be the potential product. The concept of reaction rules is also employed by enviPath²⁰, a database for predicting biodegradation mechanisms, and by MINEs²¹, a database that predicts potential biological products for mass-spectrometry applications. For a more detailed discussion of tools for reaction prediction and retrobiosynthesis, the reader is referred to Chapter 5 of this thesis.

4.1.2 The ATLAS of Biochemistry

The existing tools for predictive biochemistry are useful to answer a given research or engineering question, but they do not quantify and explore metabolic “dark matter” in a global, unbiased manner. Here, we approached the “dark matter” quest by creating an ATLAS of known and novel, predicted biochemistry from generalized enzymatic reaction rules. In the ATLAS of Biochemistry database series, abbreviated with ATLASx, we predict and reconstruct biochemical reactions within a predefined biochemical scope and store them in a database for further analysis. The aim of the ATLASx project is twofold: On one hand, we want to answer the fundamental question of unknown and engineerable biochemistries, and on the other hand, we want to provide a useful resource for synthetic biology and metabolic engineering applications. To be useful for the latter, our databases are connected to a user-friendly web interface allowing easy access to the hypothetical biochemical networks. A powerful pathway search tool further allows to answer specific research and engineering questions by searching for biotransformation routes between a source and a target compound. Thanks to the direct visualization of the pathways, the results can be manually inspected and evaluated without requiring any advanced computational skills.

The original ATLAS of Biochemistry has been created by applying the complete set of BNICE.ch reaction rules to all of the metabolites stored in the Kyoto Encyclopedia of Genes and Genomes (KEGG)²². The approach resulted in the generation of more than 130,000 novel reactions and the integration of almost 4,000 orphan KEGG compounds into at least one novel enzymatic reaction (Figure 4.1). We validated the predictive power of ATLAS by showing that more than half of the KEGG reactions added in 2015 were already part of ATLAS based on the KEGG version of 2014. More recently, Yang *et al.* experimentally validated hypothetical ATLAS reactions and used them to construct novel one-carbon assimilation pathways²³. The ATLAS of Biochemistry has been published in 2016 during my first year of PhD¹, and by the end of 2019, we had provided ATLAS access to more than 90 research groups. The interest from the community in this work encouraged us to push the development of ATLAS towards unprecedented biochemical dimensions by predicting reactions beyond the compound space of KEGG. By expanding the scope of ATLAS step by step, we hope to provide a more comprehensive overview on the hypothetical potential of metabolism and to provide better and more adapted reaction and pathway predictions to metabolic engineers.

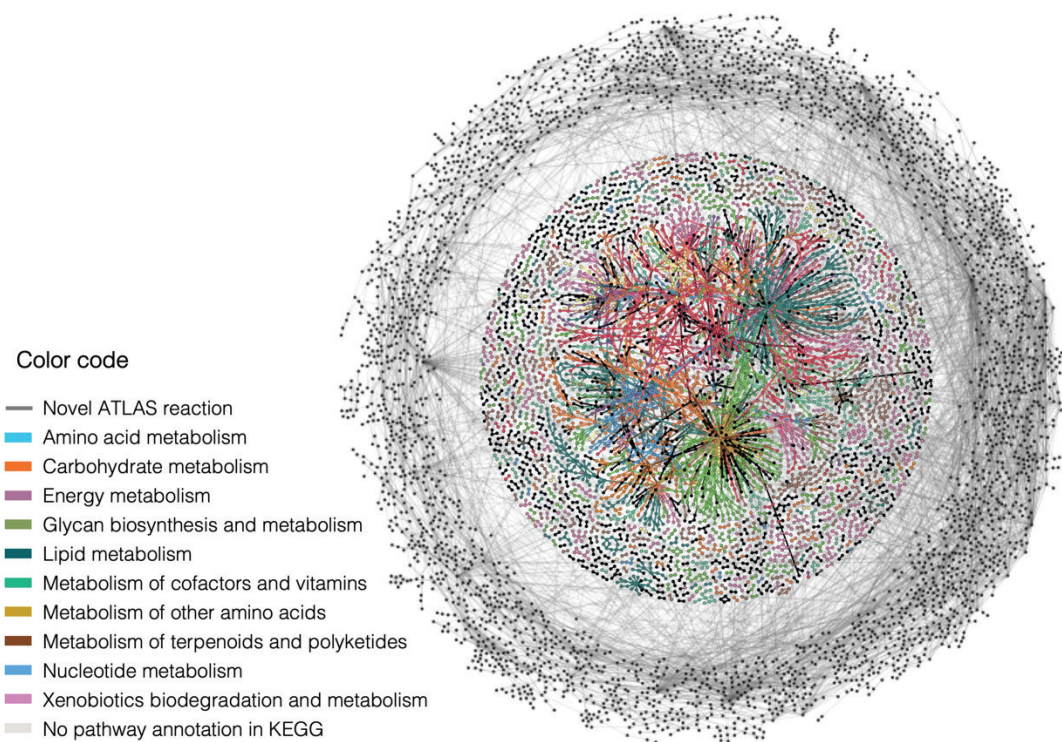


Figure 4.1: The biochemical reaction network of the ATLAS of biochemistry. Reactions are color-coded according to their pathway annotations in KEGG. The network has been drawn in the open-source graph tool Gephi²⁴.

The scope of an ATLAS project is defined by the origin of the compounds that will be connected through reaction reconstruction and prediction (Figure 4.2). In the original ATLAS for example, the scope was defined by the compound data available in KEGG. In the following, we will discuss the extension of the ATLAS scope to the biological and bioactive space in the bioATLAS project, and we will get an idea of how its extension to the chemical space, chemATLAS, will look like. Finally, we will evaluate the potential and challenges of predicting novel compounds that are not known to any biological or chemical database, and therefore part of the hypothetical compound space.

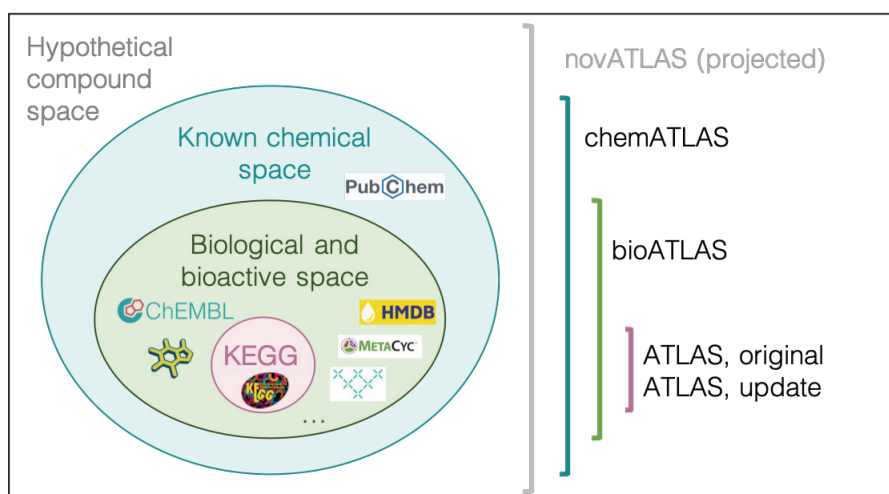


Figure 4.2: Defining the scope of different compound spaces and their associated ATLAS projects.

4.2 The ATLAS workflow

The ATLAS of Biochemistry, published in ACS Synthetic biology in 2016, was developed with Dr. Noushin Hadadi (project lead, manuscript) and Adrian Shajkofci (website development). Pipeline development, reaction reconstruction, reaction generation and data analysis were done by the author. This section only presents the ATLAS workflow, omitting any results published in the article.

Creating an ATLAS for a given biological or chemical scope requires several working steps involving different cheminformatic tools and analysis methods. The overall ATLAS workflow, discussed in the following, has been established for the original ATLAS of Biochemistry. The technical pipeline has evolved over time, but the concept stayed the same (Figure 4.3): The minimal input is a database of compounds, defining the scope of the ATLAS project, and a set of generalized reaction rules, defining the biochemical reaction mechanisms considered. Before starting the workflow, a quality check is necessary for the collected data (see 4.2.1, Database curation). If a reaction database is available for the chosen scope, as it is the case for biological databases, it can be used as a source of biochemical knowledge for defining reaction rules and later to cross-validated predicted reactions (see 4.2.2, Reconstruction of known reactions). The first step of the workflow consists of applying all of the generalized reaction rules on all of the compounds in the database using BNICE.ch, thus creating all potential reactions between the compounds (see 4.2.3, Prediction of novel reactions). If a reaction database is available, BNICE.ch will look up the generated reactions in the database and label them as known or novel (predicted). In a second step, the generated reactions are annotated with Gibbs free energy of reaction estimated by the Group Contribution Method²⁵ (GCM) and, for novel reactions, with putative enzymes using the enzyme prediction tool BridgIT²⁶ (see 4.2.4, Reaction annotation and analysis). Finally, the annotated reactions, known and predicted, are collected in a database and made available via an interactive web interface (see 4.2.5, ATLAS web interface).

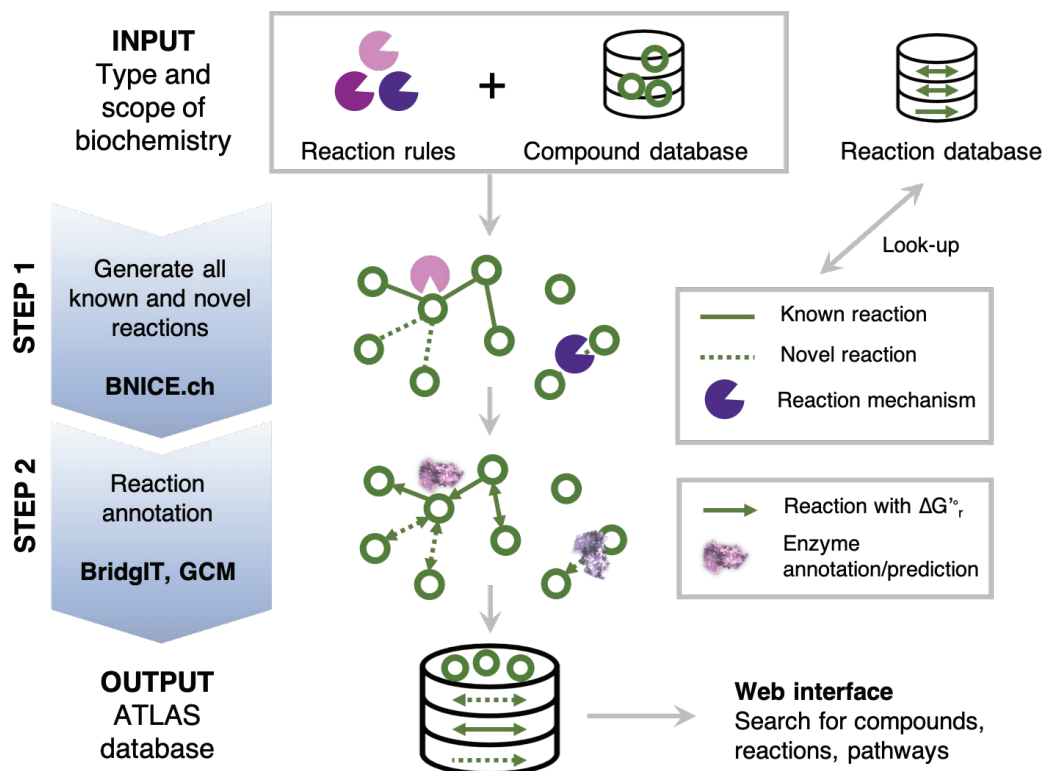


Figure 4.3: Overview on the general ATLAS workflow

4.2.1 Database curation

Each database has its own standards regarding the representation, quality and annotation of compounds and reactions. As a consequence, we first need to filter and curate the collected data before applying feeding it to BNICE.ch. We start with filtering the compounds, and then analyze the reactions based on the compounds filtering and other criteria.

To ensure that all the compounds can be read by BNICE.ch we remove compounds with undefined molecular structures, such as polymers with an unknown number of repetitions, generic compounds or entries with more than one disconnected structure, *e.g.* salts. Some compounds have R-groups, which can be handled BNICE.ch and therefore are not removed. If several databases are used in an application scope, the data needs to be unified and duplicates should be removed based on a unique identifier. It is important to note at this point that we do not treat different stereoisomers separately, but merge them into a single compound entry. The reason for this is that the downstream tool BNICE.ch was configured to not differentiate between distinct stereoisomers, in order to allow a broader prediction range of the generalized reaction rules.

In case the scope of the project is biological, reactions are collected and filtered as well. Since the reactions are not a direct input to BNICE.ch, but used for cross-checking, comparison and annotation, their processing is slightly different. Instead of removing reactions from the collection, we divide them into three sets of different quality. To be part of the top-quality set, the reaction should only involve compounds with defined molecular

structure, be elementally balanced, have a known reaction mechanism as well as a catalyzing enzyme assigned. This set can later be used to assess the BNICE.ch reaction coverage. For reactions without any enzyme assigned (*i.e.*, orphan reactions) that still contain at least one clearly defined molecular structure on each side, BNICE.ch might still find a correct reaction mechanism in its collections of reaction rules even if the cofactors are unknown or unbalanced. The remaining low-quality reactions are not used in the reaction reconstruction, but they can later be used to assess the quality of the database.

4.2.2 Reconstruction of known reactions

To assess the coverage of the collected metabolic reactions by BNICE.ch reaction rules, we try to assign each reaction with a reaction rule, and therefore with a reaction mechanism. To do this, we apply all the rules on the participating compounds of a reaction, and we check whether or not the reaction can be reconstructed in one or several reaction steps. We differentiate the following cases: (i) the reaction is exactly reconstructed, (ii) the main biotransformation is reconstructed using an alternative set of cofactors, (iii) the reaction is reconstructed in two consecutive reaction steps via compound that is part of the project scope, and (iv) the reaction is reconstructed in three reaction steps. This analysis results in the annotation of reactions with exact reaction mechanism, the corresponding third-level EC number coming from the reaction rule, as well as an estimated Gibbs free energy of reaction provided by GCM. The reaction reconstruction identifies reaction mechanisms for orphan reactions and proposes potential multi-step reaction sequences for reactions with unclear mechanisms. The reconstruction process also presents an opportunity to identify important reactions that are not yet reconstructed by any BNICE.ch reaction rule, and to add the missing mechanisms to the collection of reaction rules.

4.2.3 Prediction of novel reactions

To generate all possible reactions within the chosen scope of compounds, we apply each reaction rule to all of the compounds using BNICE.ch. Given a substrate compound, BNICE.ch generates all product structures that are possible according to the biotransformation encoded in the reaction rules. The products are compared against the compound database, and only products (and associated reactions) are retained that are part of the predefined scope of the project. The reactions are then checked against the reaction database to determine whether the reaction is known, biological, or novel, hypothetical, and added to the ATLAS database. Reactions predicted by BNICE.ch are by default assigned with a reaction mechanism, and therefore elementally balanced. The organization of reaction rules according to the Enzyme Commission (EC) classification further allows assigning a third-level EC number to the predicted reaction.

4.2.4 Reaction annotation and analysis

In order to evaluate the thermodynamic and biocatalytic feasibility of the predicted reactions, ATLAS provides an estimated Gibbs free energy ($\Delta G_r'^\circ$) for each reaction as well as enzyme predictions for novel reactions. The $\Delta G_r'^\circ$ is calculated from the Gibbs free energy

of formation ($\Delta G_r'^\circ$) of each of its reactants and products, which is estimated by the Group Contribution Method (GCM)²⁵ integrated in BNICE.ch. To assess the biocatalytic feasibility of the predicted reactions, we apply the computational enzyme prediction tool BridgIT²⁶ to each reaction. BridgIT compares the structure of the novel reaction to all known reactions, calculates a similarity score for each comparison and finds a the most similar known reaction that has an enzyme associated. A more detailed discussion of BridgIT and other enzyme prediction tools can be found in Chapter 5 (5.1.3, Finding enzymes for predicted reactions). The reaction with the highest similarity score and its associated enzyme is used to annotate the novel reactions, thus providing a metric for their biocatalytic feasibility. The annotation of reactions with estimated $\Delta G_r'^\circ$ and putative enzymes allows us to draw of the energy and EC distributions for known versus novel reactions, and to compare them.

Finally, knowledge of the exact reaction mechanism is used to determine reactant-product pairs for each reaction, which are necessary to construct a searchable graph representation of the ATLAS network. Reaction rules can be used to map atoms in a reaction, as demonstrated in Chapter 2. The atom maps can subsequently be used to calculate a Conserved Atom Ration (CAR) for each substrate-product pair in a reaction and to determine biologically relevant pairs, as shown in Chapter 3. For the first, KEGG-based ATLAS projects we used reactant pairs with a CAR above 0.34 to construct a searchable graph, which is used by the online pathway search application. For later ATLAS projects within a broader compound scope, we directly constructed an atom-weighted graph and matched it with an adapted pathway search, Yen's k-shortest loopless path search²⁷, for the efficient online prediction of pathways within the ATLAS network.

4.2.5 ATLAS web interface

The web interface has originally been developed and maintained until 2016 by Adrian Shajkofci. Since then, the database and the website have been maintained and extended by the author.

Visualizing the ATLAS data online via a web interface is essential to share the predictions with the scientific community. For the ATLAS versions within the KEGG scope, one can search through and download the reconstructed KEGG reactions only, or the totality of reactions in ATLAS. An integrated pathway search application further allows querying ATLAS for biochemical pathways between a given precursor and a target compound. The website is accessible at <https://lcsb-databases.epfl.ch/pathways/atlas/>, with free registration for academia upon subscription.

4.3 The update - ATLAS 2018

This Subchapter will be submitted to the journal ACS Synthetic Biology as a technical note. The work has been achieved in collaboration with Homa Mohammadi-Peyhani (reaction generation, manuscript), Anastasia Sveshnikova (enzyme prediction) and Alan Scheidegger (compilation of reactions). The author has been in charge of the project lead, the manuscript and the reconstruction of KEGG reactions.

The ATLAS of Biochemistry⁸ was created based on the biochemical knowledge available in KEGG 2015²². Since then, KEGG has added 802 new metabolites, 918 new reactions, and 633 enzymes to its collection. We took advantage of this newly available data to validate some of the reactions and enzymes predicted in 2015, and we seized the opportunity to integrate the new compounds and reactions into ATLAS to expand the space of predicted reactions. Furthermore, two main aspects of our workflow had been significantly improved in the meantime. First, the set of bidirectional reaction rules was increased from 360 to 400. Many of the new rules were created to reconstruct some more complex reactions of secondary metabolism, which are more difficult to be generalized, and hence were not considered previously. Second, BridgIT has been adapted to integrate information about the reactive site into the comparison of reactions, which greatly improved its predictive power. In the following, we discuss the updated ATLAS statistics and illustrate the improvements compared to the first version, referred to as ATLAS 2015. The latest version of ATLAS, referred to as ATLAS 2018, is available online (<http://lcsb-databases.epfl.ch/atlas/>).

4.3.1 Reconstruction of known reactions

We applied the previously described ATLAS methodology to the newly available data, and we compared the generated reactions to the metabolic reactions stored in KEGG. The KEGG database contained 18,254 compounds as of February 2018 (Table 4.1). 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 4,587 were not involved in any KEGG reaction. These disconnected compounds, called “orphan” metabolites, 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, 8,041 reactions were reconstructed with BNICE.ch reaction rules and 5,780 reactions were exactly reconstructed. Another 1,708 reactions were reconstructed using alternative cofactors, out of which 123 reactions were poorly characterized in KEGG (i.e., reaction mechanism not known, incomplete reaction). The remaining 553 reactions were reconstructed in two (408 reactions) or three (145 reactions) reaction steps.

A total of 2,712 KEGG reactions were not reconstructed with BNICE.ch. First, 1,544 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 860 reactions were not reconstructed because their reaction mechanisms are very specific and hence not readily generalizable.

4.3.2 Validation of novel ATLAS reactions

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 996 predicted reactions in ATLAS were not actually novel, but known to at least one of the repositories Brenda, Reactome, HMR, MetaCyc, 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.

4.3.3 ATLAS 2018 statistics

ATLAS 2018, based on KEGG 2018, now has 149,052 reactions, out of which 5,780 are known to KEGG. Compared to 2015, we added 385 known and 11,173 novel reactions. Thanks to predicted reactions, ATLAS now integrates 4,587 out of 9,857 orphan KEGG metabolites, meaning those that do not have a known catalyzing enzyme.

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 4.1). 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, all of them with a higher score than in the previous version (Figure 4.4).

Table 4.1: Overview on compound, reaction and enzyme statistics in KEGG and ATLAS

		ATLAS 2015	ATLAS 2018	Change
KEGG compounds	Total number of compounds	17,450	18,254	+ 5%
	Filtered compounds (<i>fc</i>)	16,798	17,255	
	Orphan KEGG compounds (<i>okc</i>)	9,371 (56% of <i>fc</i>)	9,857 (57% of <i>fc</i>)	
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	8,041	+20%
	Exact coverage	5,270	5,780	
	Alternative cofactor usage	916	1,708	
	2-step reconstruction	387	408	
	3-step reconstruction	78	145	
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 <i>okc</i>)	4,587 (47% of <i>okc</i>)	

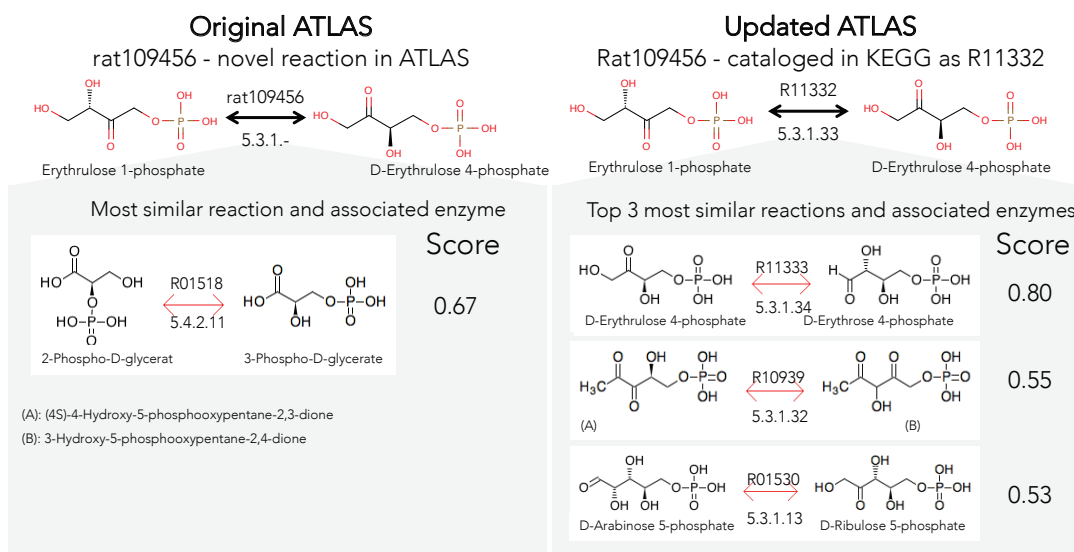


Figure 4.4: The reaction with ATLAS identifier rat109456 is an example of a reaction that was novel in ATLAS 2015 and that is now cataloged in KEGG. (left) rat109456 along with its most similar reaction and candidate enzyme, predicted by an earlier version of BridgIT to calculate structural reaction similarities. (right) rat109456 in ATLAS 2018 is now cataloged in KEGG as R11332 with EC 5.3.1.33. Two alternative enzyme candidates are proposed by the updated version of BridgIT.

We have updated the ATLAS of Biochemistry to integrate new biochemical data from KEGG 2018 using an updated set of generalized reaction rules and by employing an improved version of BridgIT to enhance the enzyme predictions for novel reactions. This study demonstrates the dynamic nature of biochemical knowledge and highlights the need for continuous updates of database-dependent applications. In particular, we showed that integrating databases other than KEGG into the ATLAS workflow adds value to the prediction of reaction and enzyme.

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

The results presented in this Subchapter have been obtained in collaboration with Homa Mohammadi-Peyhani (data collection, reaction generation, pipeline and database development), Anastasia Sveshnikova (reactive site analysis) and Victor Viterbo (curation of reaction databases). The author has been in charge of the project lead, the manuscript, as well as pipeline and website development.

One major drawback of ATLAS is, however, its limitation to KEGG compounds. We estimated that integrating molecular structures from different biological, biochemical and chemical database would greatly enhance the application range and the predictive power of the database: For example, many drugs and plant natural products with undefined or putative biological function were not included in KEGG compounds scope. Furthermore, we showed that in many cases, compounds from databases other than KEGG can help to reconstruct multi-step metabolic reactions, or to bridge broken biosynthesis routes towards secondary metabolites. 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.

To achieve this this, we decided to expand the ATLAS scope to all known biological and biochemical compounds (Figure 4.5). We first collected over 60,000 biochemical reactions from ten publicly available databases, and over 1.5 million biological and bioactive compounds from eight databases, and we stored the data in a database called bioDB. The compounds in bioDB form the *biological and bioactive compound space*. We also collected all compounds from the chemical database PubChem, representing the *chemical space*. We then applied the ATLAS workflow, as described in Subchapter 4.2, to the biological and bioactive compound space using an updated set of 447 bidirectional reaction rules. We generated all biochemically possible reactions whose products were part of the biological and bioactive compound space, stored in the *bioATLAS* database. Products that were belonging to the chemical space were stored as *chemATLAS*. The result is a database of more than 1.7 million known and novel biochemical reactions connecting one million biological and bioactive compounds. Due to the increased size of the compound space, some of the steps in the workflow had to be adapted and will be discussed in the next subsection.

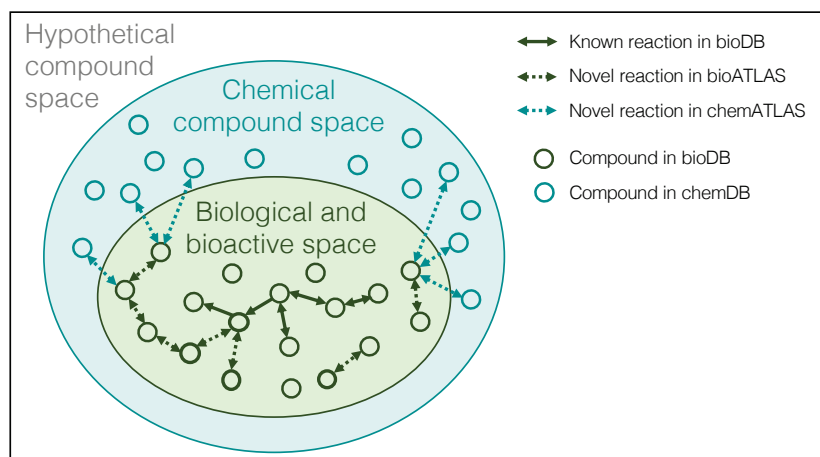


Figure 4.5: Different types of reactions in ATLAS derived from of the biological and bioactive compound space.

4.4.1 New tools & methods

The first versions of ATLAS were small enough to be stored and analyzed in a text-based format. Based on our experience with the KEGG ATLAS, we projected to exceed an estimated one million reactions for bioATLAS, which would make text-based analysis difficult and inefficient. To be able to efficiently store, retrieve and analyze the increased amount of data, we created an SQL-based database hosted on our in-house server. The database architecture has been developed by my colleague Homa Mohammadi-Peyhani and is not further discussed here. The new database format required changes in the BNICE.ch framework, which closely interacts with the database, as well as in the web interface that displays database content. One important change is the switch from a data-heavy, PNG-based visualization of compounds to the SMILES viewer developed by Probst and Reymond²⁸. Furthermore, we took advantage of the graph extraction method presented in Chapter 3 to represent the generated data as an atom-weighted graph and to perform pathway search within the hypothetical biochemical network.

Given the increased number of compounds to be analyzed, we divided the reaction generation process into two consecutive steps. The first step consists of screening all the compounds in the database for reactive sites. For this, we checked if the reactive site description of a given reaction rule matches any substructure in the compounds database. We kept the information on which reaction rules could recognize which compounds, and we analyzed this first result separately. In a second step, we applied each reaction rule to the compounds known to be recognized by the rule in question to generate all possible reactions.

Finally, the increased size of the generated biochemical networks required powerful graph search and analysis methods. For this, we employed the tools developed in Chapter 3 that are based on the Python toolbox NetworkX. For the efficient calculation of network diameters we used the Python version of the SNAP toolbox²⁹.

4.4.2 Collecting biochemical data

To start with, we selected the publicly available biochemical databases that match the biological and bioactive scope. All compound entries were collected from MetaCyc, Model SEED, KEGG compound and drug databases, Drugbank, ChEBI, HMDB, MetaNetX, and ChEMBL (Table 4.2). From the collected compounds, only entries associated to a molecular structure were imported to our database. Next, the imported compounds were unified based on their 2D canonical SMILES, and annotations from different databases were merged into one compound entry in the database, resulting in 1,698,524 unique compounds. As a result, different tautomers, stereoisomers and charged states of a same compound were merged into one compound entry in our database. However, not all of these compounds could be used as a direct input for the ATLAS workflow. Compound entries that describe more than one disjoint molecular structure, *e.g.* salts, were excluded from the bioATLAS scope. As a result, the input for the ATLAS workflow summed up to 1,500,222 biological and bioactive compounds.

Table 4.2: Compounds from different sources imported to bioDB

	Databases	Description	Collected	Imported	Unique
BIOLOGICAL	MetaCyc	Metabolites found in sequenced organisms	15,819	14,828	13,499
	Model SEED	Metabolites from KEGG and GEMs	33,995	20,665	18,436
	KEGG Compound	Compounds and biopolymers relevant to biology	18,625	17,397	15,869
BIOACTIVE	KEGG Drug	Approved drugs in Japan, USA & Europe	11,140	7,766	7,765
	Drugbank	Approved & discovery-phase drugs	8,350	6,279	4,217
	ChEBI	Chemical Entities of Biological Interest	56,530	32,691	32,352
	HMDB	Small metabolites found in the human body	228,017	177,096	100,031
	MetaNetX	Metabolites found in GEMs (excl. lipids)	200,132	183,788	128,308
	ChEMBL	Bioactive compounds	1,727,112	1,595,615	1,522,924
Total bioDB			2,297,709	2,056,125	1,698,524
bioATLAS input					1,500,222
Total chemDB					77,934,143

We also imported all chemical compounds that can be found in PubChem³⁰ for further comparison, although they were not used as an input for the ATLAS workflow in this study. Any known compounds that can be found in the chemical database PubChem, but that is not part of the bioDB, was assigned to the *chemical space*. We are aware that this classification is somewhat artificial, since compounds from the chemical space may actually be of biological origin, yet they are not present in any biological or bioactive database. We will see later that this artificial classification hides an opportunity to re-assign compounds of the chemical space to the biological compound space. The unfiltered chemical space counts a total of 77,934,143 unique compounds and is referred to as chemDB.

Reactions were collected from KEGG, BRENDA³¹, Rhea³², BiGG³³, SEED³⁴, MetaNetX³⁵, MetaCyc³⁶, HMDB³⁷, Reactome³⁸ and BKMS-react³⁹ (Table 4.3). After filtering out reactions containing undefined molecular structures (*e.g.* polymers, proteins) and reactions that were not elementally balanced, 29,637 reactions were left. Out of these, we could assign BNICE.ch reaction mechanisms to 15,474 of them.

Table 4.3: Overview on collected reactions from different databases and their reconstruction with BNICE.ch reaction rules

Data-bases	Col-lected	Imported	Unique	Non-orphan	Elementally balanced	Reaction mechanism*
HMDB	8,182	5,108	4,380	3,417	3,177	1,276
MetaCyc	16,052	15,438	12,726	9,614	7,879	6,177
KEGG	10,829	10,685	10,179	9,667	9,010	7,230
Meta-NetX	42,182	40,767	25,647	14,194	12,733	6,944
Reactome	1,872	1,568	814	342	406	213
Rhea	20,770	19,325	13,114	10,808	10,401	6,045
Model SEED	44,031	44,010	28,332	9,816	14,290	6,773
BKMS	31,740	18,139	18,139	15,493	10,556	7,742
BiGG	28,299	16,581	8,681	3,874	3,445	947
BRENDA	31,741	9,214	7,044	6,629	6,825	4,310
Total	235,698	180,835	61,234	30,376	29,637	15,474

*Reaction mechanism present in BNICE.ch reaction rule database

To summarize, we unified biochemical reactions and compounds from a total of fifteen different sources into one database, named bioDB. bioDB holds 1,500,222 unique biological or bioactive compounds, and 61,308 unique biochemical reactions. Around one fourth of the reactions are associated to a reaction mechanism as encoded in the BNICE.ch reaction rules.

4.4.3 Analysis of reactive sites

As a first analysis for the collected biological and bioactive compounds, we aimed to determine their biochemical reactivity of the by screening them for reactive sites. To achieve this, we applied the reactive site recognition encoded in the BNICE.ch reaction rules to all of the compounds in bioDB. As a result, each compound was assigned a list of reaction rules that can recognize one or more reactive sites on the molecular structure. The number of reaction rules assigned to a compound is as an indicator for the diversity of functional groups, or the biochemical versatility, of the molecule.

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. By looking at the distribution of reactive sites versus number of carbon atoms in the molecule, we found that most of the compounds (87%) had between 50 and 200 reaction rules assigned, within a range from five to twenty carbon atoms (Figure 4.6A).

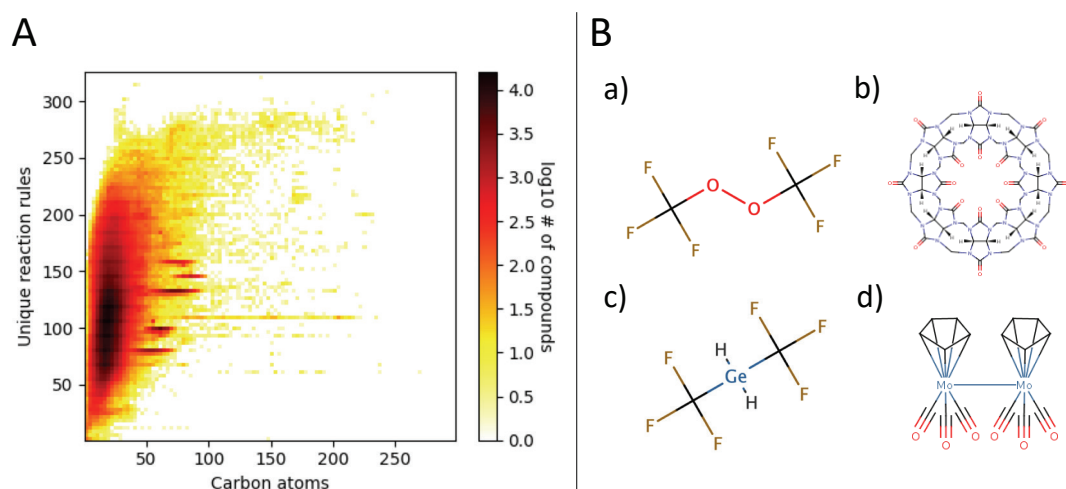


Figure 4.6: (A) Heatmap showing the distribution of compounds as a function of their number of carbon atoms versus the number of reaction rules assigned to them. The color indicates the number of compounds on a logarithmic scale. (B) Four compounds for which BNICE.ch could not find any reactive site. a) Bis(trifluoromethyl)peroxide (BTP), b) cucurbit[8]uril, c) Bis(trifluoromethyl)germane, d) bis[tricarbonyl(η5-cyclopentadienyl)molybdenum](Mo—Mo).

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). 16 compounds contained only one carbon atom that was not accessible to metabolism (*e.g.*, CFe8S9). The remaining four compounds were found to be chemically synthesized molecules with medical or research applications (Figure 4.6B). Even though these

compounds do not seem to have the chemical capacity to participate in any biochemical reaction, their presence in biological databases can still be justified through their interaction with living organisms.

4.4.4 Prediction of novel reactions

At the time of writing, the reaction prediction process was still ongoing. All of the results discussed in the following represent an intermediary snapshot of work in progress (26 December 1019). Daily updated database and network statistics can be found online under <https://lcsb-databases.epfl.ch/Atlas2/Statistics>.

Next, we expanded bioATLAS from the compound space in bioDB by predicting novel, hypothetical reactions from known biological and bioactive compounds. To achieve this, we applied the 477 bidirectional reaction rules on the 1,498,307 in bioDB that were assigned at least one reaction rule in the previous step of the workflow. Reactions whose products were part of the biological and bioactive compounds space were stored to bioATLAS, and reactions whose products that were only part of the chemical compound space were stored to chemATLAS. We predicted a total of 5,389,453 novel reactions from biological and bioactive compounds. 1,711,285 (32%) out them occurred exclusively between biological and bioactive compounds, and the remaining 3,678,168 reactions involved at least one compound from the chemical space. In terms of compounds, bioATLAS integrates almost two thirds (906,979 out of 1,500,222) of the compounds fed to the ATLAS workflow. One of the objectives of the ATLAS method is to integrate orphan compounds into the biochemical reaction space. The bioDB counts 1,492,594 orphan compounds that are not involved in any known reaction, even though they are labeled as biological or bioactive molecules. Interestingly, 60% (899,351) out of these orphan compounds could be integrated into at least one novel ATLAS reaction.

4.4.5 Network analysis

To analyze the properties of the newly created biochemical network, we explored bioDB, bioATLAS and chemATLAS from a graph-theoretical point of view. Graph theory has been repeatedly employed to analyze metabolic networks, by representing compounds as nodes and biochemical transformations as edges. However, most methods either are based on manually derived reactant-product pairs (*e.g.*, KEGG RPAIR network⁴⁰), or they define a set of cofactors to be excluded from the analysis to avoid the generation of hubs by currency metabolites. Here, we rely on the method proposed in Chapter 3, called NICEpath, that weights each substrate-product pair according the number of atoms conserved between the substrate and the product. The weight, or Conserved Atom Ratio (CAR), is then transformed into a distance between the substrate node and the product node, and the two are connected by an edge representing the biotransformation. Since a same biotransformation, or reactant-product pair, can occur in more than one reaction, each edge in the network represents all the reactions that transform the substrate to the product. The proposed method allows the construction of an atom-weighted graph that can be used to find metabolic pathways by using common graph search algorithms. Moreover, the atom-weighted

graph can be analyzed to derive global properties of the biochemical networks. To calculate the weights on each pair, NICEpath requires that each reaction is annotated with a reaction mechanism that allows the calculation of the atom conservation between the substrate and the product. This condition is met for all bioDB reactions with assigned reaction rules, and for all predicted reactions in bioATLAS and chemATLAS.

We constructed three networks with different biochemical scopes: The 15,474 mechanistically curated reactions in bioDB translated into 15,142 weighted edges connecting 7,628 bioDB compounds. The bioATLAS network connects 906,979 compounds in 1,711,285 reactions, represented by 2,487,983 edges, and chemATLAS connects 1,924,960 compounds in 5,389,453 reactions, represented by 5,600,259 edges (Table 4.4). For many types of network analysis however, an unweighted graph is required. Since we know from previous studies presented in Chapter 3, substrate-product pairs with a very low degree of atom conservation may not be biologically relevant. We could show that a cut-off of 0.34 in the conserved atom ratio predicts best predicts the manually curated reactant pairs of type “main” in KEGG⁴⁰. Hence, we removed edges with a CAR below 0.34, and we removed the weights from the remaining edges. The result is a new set of unweighted graphs, which can be analyzed using standard graph analysis algorithms.

The connectivity of a biochemical reaction network can give us insights into the comprehensiveness our knowledge, and help us identifying missing biochemical links. To assess the connectivity of the graph, we counted the number of connected components, *i.e.*, disjoint graphs, or islands in the network (Figure 4.7A). We found that the total number of components increased with the network expansion from bioDB to bioATLAS to chemATLAS (Table 4.4). However, the number of components relative to the size of the network, or the number of components divided by number of nodes, decreased from 0.17 in bioDB, to 0.10 in bioATLAS, to 0.08 in chemATLAS, suggesting that the network becomes more connected by increasing the scope expansion. Next, we had a closer look at the size distribution of the components and we found that each of the networks was dominated by one big component, followed by a big number of secondary components of maximal 16 compounds involved (Figure 4.7B). While the biggest component in bioDB only includes 59% of the number of edges, this number rose to 73% in bioATLAS and reached almost 80% in chemATLAS. From this, we can conclude that integrating bioactive and chemical compounds makes the biochemical network denser. 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 decreased in bioATLAS and chemATLAS compared to bioDB, which means that the shortest connections between any two nodes has decreased on average. The network diameter was decreased from bioDB to bioATLAS, representing a densification of the network, and increased again when including chemical compounds, suggesting expansion of the network towards novel chemistry and integration of previously disconnected components.

Table 4.4: Network statistics of bioDB, bioATLAS and chemATLAS networks.

Network	Property	bioDB	bioATLAS	chemATLAS
Weighted network	Number of nodes	7,628	906,979	1,924,960
	Number of edges (CAR > 0)	15,142	2,487,983	5,600,259
Un-weighted network	Number of nodes	7,537	703,046	1,902,991
	Number of edges, unweighted (CAR > 0.34)	7,460	1,058,683	2,757,348
	Number of components (disjoint graphs)	1,259	89,926	162,920
Biggest component	Number of nodes	3,359	356,889	1,235,405
	Number of edges	4,405	769,999	2,193,872
	Percent of total number of nodes	44.57 %	50.76 %	64.92 %
	Percent of total number of edges	59.05 %	72.73 %	79.56 %
	Diameter (longest shortest path)	43	32	38
	Effective diameter (average shortest path)	12	10	10

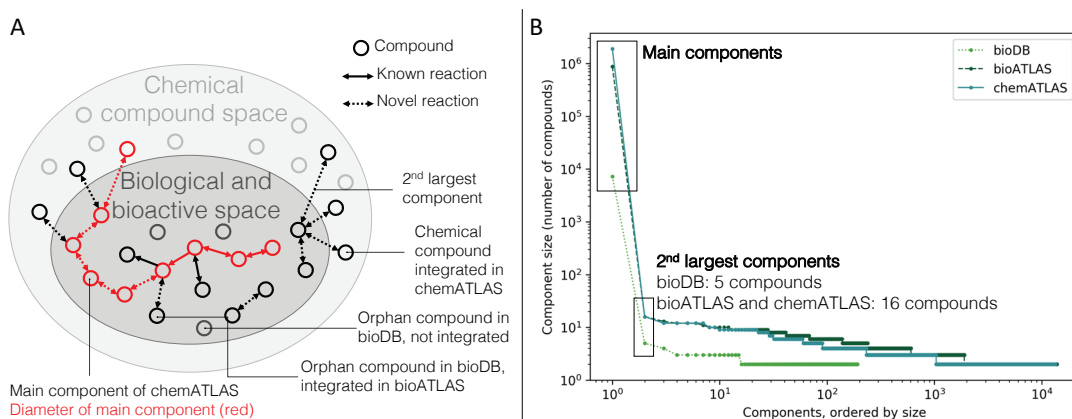


Figure 4.7: (A) Visual overview on different statistics and network properties calculated for bioDB, bioATLAS and chemATLAS. (B) Size distribution of disconnected components in the network of each of the three database scopes.

4.5 Conclusion and outlook

Based on the 1.5 million known biological and bioactive compounds in bioDB, we predicted more than 1.7 million biochemically possible biotransformations between biological and bioactive compounds using 477 generalized reaction rules, and we stored them in bioATLAS. We further predicted more than 3.6 million reactions involving compounds from the chemical space, resulting in a total of almost 5.4 million reactions in our database. From this new wealth of information, we extract insightful numbers on the connectivity and reactivity of biologically relevant molecules. Finally, we provide public access to our database through an online search interface including a powerful pathway search algorithm, which can be used for the design of novel metabolic pathways. Daily updated database statistics as well as a preliminary version of the search interface can be accessed at <https://lcsb-databases.epfl.ch/Atlas2>.

While the ATLAS expansion of biologically relevant compounds is complete, only a fraction (0.025%) of the chemical space has been integrated into biochemical reactions so far. Further work will include the systematic screening of the 75 million PubChem compounds for potentially biochemically active structures, and the prediction of biochemical reaction between them. Currently, the bioATLAS reaction generation is in its final phase, and the bio-derived part of chemATLAS is already available (Table 4.5).

Table 4.5: Overview on the past, current and projected development of the ATLASx databases

Database	Compound scope	# Compounds screened	# Compounds integrated	# Reactions (total)	# Known reactions	# Reaction rules	State of development
ATLAS 2015	KEGG 2015	16,798	10,362	137,877	5,370	361	Published
ATLAS 2018	KEGG 2018	17,255	10,939	149,052	5,780	400	Submitted
bioATLAS	All biological and bioactive databases (2019)	1,500,222	906,898	1,711,285	15,474	447	Work in progress
chemATLAS	All chemical databases	77,934,143	1,912,769	5,389,453	15,474	447	Work in progress
novATLAS	BNICE.ch predicted novel compounds	-	-	-	-	-	Projected

From our experience working with different classes of compounds, we learned that for many of them, although naturally produced by organisms, are not recorded in any database. A typical example is plant secondary metabolism, where the promiscuous activity of so-called “decoration” enzymes (*e.g.*, methyltransferases) may lead to the production of small amounts of diverse derivatives of the known secondary metabolites. Another case are

short-lived intermediate species that are unstable or quickly consumed by enzymes downstream in the biosynthesis path. Even though the chemical structures of these intermediates have never been studied, their chemistry is important in the characterization and design of biosynthetic pathways. While BNICE.ch predicts these potential novel derivatives, they are usually excluded from the solution space in the BNICE.ch network generation because they easily lead to combinatorial explosions in the number of potential products with each iteration. In an ATLAS-type of approach however, novel compounds could be integrated in a more targeted fashion: By predicting novel compounds around each compound, we could avoid combinatorial explosion by only keeping novel structures that connect to two known compounds (*i.e.*, a node degree of two or more), or by applying machine learning techniques that predict the biochemical feasibility of the novel structures to filter out compounds with a low probability to exist in biological conditions. Exploring the space of hypothetical biochemical structures will eventually help us to predict compounds that are difficult to be detected experimentally, such as hypothetical metabolic intermediates or secondary metabolites produced at very low concentrations. The creation of an ATLAS reaching out to the chemical space of novel compounds, novATLAS, will further help to map dark matter in metabolism.

To conclude this chapter, the ATLASx project is a dynamic, continuously evolving effort to integrate the scattered knowledge of metabolism into an overall hypothetical biochemical reaction network. The future discovery of new metabolic structures and enzymatic reaction mechanisms will continue to feed the ATLAS workflow and, consequently, to improve our predictions for metabolic engineering applications.

References

1. Hadadi, N., Hafner, J., Shajkofci, A., Zisaki, A. & Hatzimanikatis, V. ATLAS of Biochemistry: A Repository of All Possible Biochemical Reactions for Synthetic Biology and Metabolic Engineering Studies. *ACS Synth. Biol.* (2016). doi:10.1021/acssynbio.6b00054
2. Notebaart, R. A., Kintsjes, B., Feist, A. M. & Papp, B. Underground metabolism: network-level perspective and biotechnological potential. *Curr. Opin. Biotechnol.* **49**, 108–114 (2018).
3. Rosenberg, J. & Commichau, F. M. Harnessing Underground Metabolism for Pathway Development. *Trends Biotechnol.* **37**, 29–37 (2019).
4. Lerma-Ortiz, C. *et al.* 'Nothing of chemistry disappears in biology': the Top 30 damage-prone endogenous metabolites. *Biochem. Soc. Trans.* **44**, 961–971 (2016).
5. Oliver, S. G. From DNA sequence to biological function. *Nature* **379**, 597–600 (1996).
6. Galperin, M. Y. & Koonin, E. V. From complete genome sequence to 'complete' understanding? *Trends Biotechnol.* **28**, 398–406 (2010).
7. Keseler, I. M. *et al.* The EcoCyc database: reflecting new knowledge about *Escherichia coli* K-12. *Nucleic Acids Res.* **45**, D543–D550 (2017).
8. Hadadi, N., Hafner, J., Shajkofci, A., Zisaki, A. & Hatzimanikatis, V. ATLAS of Biochemistry: A Repository of All Possible Biochemical Reactions for Synthetic Biology and Metabolic Engineering Studies. *ACS Synth. Biol.* **5**, 1155–1166 (2016).
9. Bachmann, B. O. Biosynthesis: Is it time to go retro? *Nat. Chem. Biol.* **6**, 390–393 (2010).
10. Hadadi, N. & Hatzimanikatis, V. Design of computational retrobiosynthesis tools for the design of de novo synthetic pathways. *Curr. Opin. Chem. Biol.* **28**, 99–104 (2015).
11. Wang, L., Ng, C. Y., Dash, S. & Maranas, C. D. Exploring the combinatorial space of complete pathways to chemicals. *Biochem. Soc. Trans.* **46**, 513–522 (2018).
12. Lin, G.-M., Warden-Rothman, R. & Voigt, C. A. Retrosynthetic design of metabolic pathways to chemicals not found in nature. *Curr. Opin. Syst. Biol.* (2019). doi:10.1016/J.COISB.2019.04.004
13. Jeffryes, J. G., Seaver, S. M. D., Faria, J. P. & Henry, C. S. A pathway for every product? Tools to discover and design plant metabolism. *Plant Sci.* (2018). doi:10.1016/J.PLANTSCI.2018.03.025
14. Hatzimanikatis, V. *et al.* Exploring the diversity of complex metabolic networks. *Bioinformatics* **21**, 1603–1609 (2005).
15. Tokić, M. *et al.* Discovery and Evaluation of Biosynthetic Pathways for the Production of Five Methyl Ethyl Ketone Precursors. *ACS Synth. Biol.* acssynbio.8b00049 (2018).

doi:10.1021/acssynbio.8b00049

16. Delépine, B., Duigou, T., Carbonell, P. & Faulon, J.-L. RetroPath2.0: A retrosynthesis workflow for metabolic engineers. *Metab. Eng.* **45**, 158–170 (2018).
17. Koch, M., Duigou, T. & Faulon, J.-L. Reinforcement Learning for Bio-Retrosynthesis. *bioRxiv* 800474 (2019). doi:10.1101/800474
18. Kumar, A., Wang, L., Ng, C. Y. & Maranas, C. D. Pathway design using de novo steps through uncharted biochemical spaces. *Nat. Commun.* **9**, 184 (2018).
19. Sivakumar, T. V., Giri, V., Park, J. H., Kim, T. Y. & Bhaduri, A. ReactPRED: A tool to predict and analyze biochemical reactions. *Bioinformatics* btw491 (2016). doi:10.1093/bioinformatics/btw491
20. Wicker, J. *et al.* enviPath – The environmental contaminant biotransformation pathway resource. *Nucleic Acids Res.* gkv1229 (2015). doi:10.1093/nar/gkv1229
21. Jeffries, J. G. *et al.* MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. *J. Cheminform.* **7**, 44 (2015).
22. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
23. Yang, X. *et al.* Systematic design and in vitro validation of novel one-carbon assimilation pathways. *Metab. Eng.* **56**, 142–153 (2019).
24. Bastian, M., Heymann, S. & Jacomy, M. *Gephi: An Open Source Software for Exploring and Manipulating Networks Visualization and Exploration of Large Graphs*.
25. Jankowski, M. D., Henry, C. S., Broadbelt, L. J. & Hatzimanikatis, V. Group Contribution Method for Thermodynamic Analysis of Complex Metabolic Networks. *Biophys. J.* **95**, 1487–1499 (2008).
26. Hadadi, N., MohammadiPeyhani, H., Miskovic, L., Seijo, M. & Hatzimanikatis, V. Enzyme annotation for orphan and novel reactions using knowledge of substrate reactive sites. *Proc. Natl. Acad. Sci. U. S. A.* 201818877 (2019). doi:10.1073/pnas.1818877116
27. Yen, J. Y. Finding the K Shortest Loopless Paths in a Network. *Manage. Sci.* **17**, 712–716 (1971).
28. Probst, D. & Reymond, J.-L. SmilesDrawer: Parsing and Drawing SMILES-Encoded Molecular Structures Using Client-Side JavaScript. *J. Chem. Inf. Model.* **58**, 1–7 (2018).
29. Leskovec, J. & Sosič, R. SNAP. *ACM Trans. Intell. Syst. Technol.* **8**, 1–20 (2016).
30. Kim, S. *et al.* PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* **47**, D1102–D1109 (2019).
31. Schomburg, I. *et al.* BRENDA: a resource for enzyme data and metabolic information. *Trends Biochem. Sci.* **27**, 54–56 (2002).

32. Morgat, A. *et al.* Updates in Rhea--a manually curated resource of biochemical reactions. *Nucleic Acids Res.* **43**, D459-64 (2015).
33. Schellenberger, J., Park, J. O., Conrad, T. M. & Palsson, B. Ø. BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics* **11**, 213 (2010).
34. Aziz, R. K. *et al.* SEED Servers: High-Performance Access to the SEED Genomes, Annotations, and Metabolic Models. *PLoS One* **7**, e48053 (2012).
35. Moretti, S. *et al.* MetaNetX/MNXref – reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. *Nucleic Acids Res.* **44**, D523–D526 (2016).
36. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res.* **46**, D633–D639 (2018).
37. Wishart, D. S. *et al.* HMDB: the Human Metabolome Database. *Nucleic Acids Res.* **35**, D521-6 (2007).
38. Croft, D. *et al.* Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* **39**, D691–D697 (2011).
39. Jeske, L., Placzek, S., Schomburg, I., Chang, A. & Schomburg, D. BRENDA in 2019: a European ELIXIR core data resource. *Nucleic Acids Res.* **47**, D542–D549 (2019).
40. Shimizu, Y., Hattori, M., Goto, S. & Kanehisa, M. Generalized Reaction Patterns for Prediction of Unknown Enzymatic Reactions. *Genome Informatics* **20**, 149–158 (2008).

Chapter 5 Applications: Predicting biotransformations with cheminformatic tools

The tools, methods and databases described in the previous chapters have been applied to solve a variety of research questions and metabolic engineering problems. Depending on the scientific questions, different computational tools and resources have been combined to provide an optimal, adapted answer. At the very core of these applications is always BNICE.ch, originally designed to solve retrobiosynthesis problems. In this chapter, we will first discuss retrobiosynthesis in general (Subchapter 5.1), followed by an illustration of the theory of a typical retrobiosynthesis problem (Subchapter 5.2). We further introduce two alternative applications of BNICE.ch: The first one is the prediction of potential engineering targets and their associated bioproduction pathways that can be derived from a given metabolic pathways (Subchapter 5.3), and the second study examines the ability of BNICE.ch to predict the biodegradability of xenobiotic chemicals (Subchapter 5.4).

5.1 Retrobiosynthesis

Retrobiosynthesis is a well-established method to discover potential bioproduction routes for chemicals of industrial interest. The underlying principle of “walking back” from a target compound, reaction step by reaction step, until a suitable precursor is reached, has first been explored by organic chemists^{1–3}. In chemical retrosynthesis, reversed synthetic reaction rules are applied iteratively on a target compound. The result of this procedure are possible synthetic pathways, connecting the target molecule for synthesis back to cheaper commodity chemicals. The approach has been adapted to biosynthetic pathways, giving rise to a range of retrobiosynthesis tools that try to connect the target compounds to biologically available precursor compounds, but only a few of them are subject to continuous development. The most popular, continuously maintained retrobiosynthesis tools are BNICE.ch^{4,5}, RetroPath^{6,7} and novoStoic^{8,9}. Recent reviews provide a critical overview of these tools and discusses in detail their potential as well as their limitations^{10–14}.

Compared to their counterparts from chemical synthesis, retrobiosynthesis methods have an additional set of challenges to solve. Aspects such as the availability of the precursor in the host organism, availability of cofactors, thermodynamic feasibility in ambient conditions, the toxicity of intermediates and the availability of enzymatic catalysts should be

considered before trying to implement a predicted pathway into the host. Different tools considering different aspects are summarized in Table 5.1.

This Subchapter describes the state-of-the-art workflow of retrobiosynthesis based on the methods described in the previous chapters. The workflow is divided into four main steps (Figure 5.1): (i) The generation of a hypothetical biochemical reaction network around the target, (ii) the search for metabolic pathways that connect the target to potential precursor compounds produced by the chassis organism, (iii) stoichiometric and thermodynamic feasibility of the proposed pathway within the genome-scale model of the chassis organism, and (iv) finding suitable enzymes for each known and predicted reaction step in the pathway. For each part of the workflow, we list alternative method that can be used instead of the BNICE.ch-based tools.

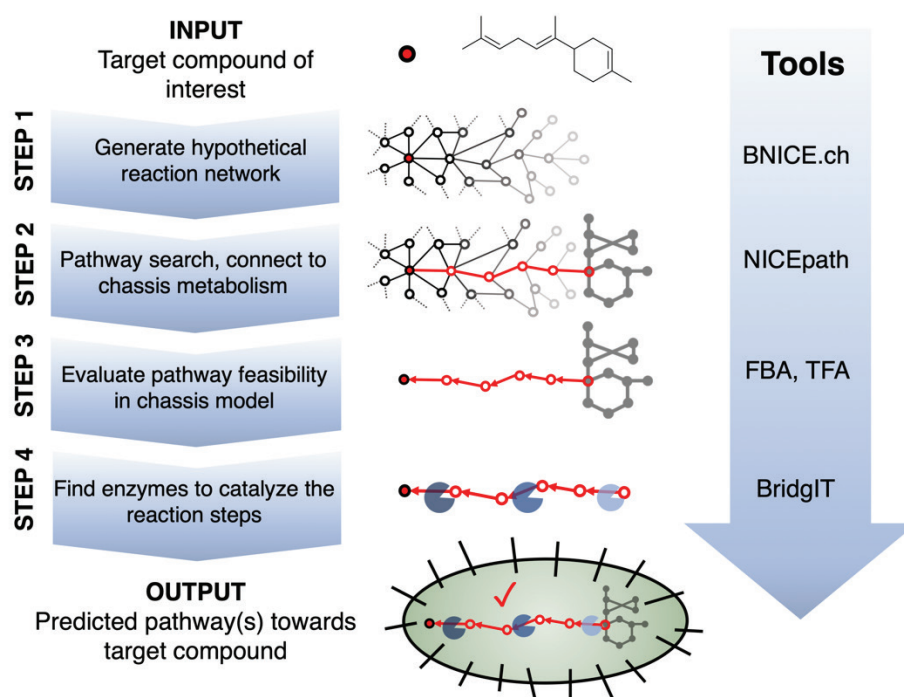


Figure 5.1: Schematic of BNICE.ch-based retrobiosynthesis workflow. A hypothetical biochemical network is expanded around the target compound (red dot). Pathways connecting the host metabolism to the target (red path) are retrieved and evaluated for stoichiometric, thermodynamic and enzymatic feasibility.

Table 5.1: Available retrobiosynthesis tools and their characteristics (adapted and updated from Hadadi and Hatzimanikatis¹⁰).

Tool	Generalized reaction rules	Gibbs free energy of reaction	Network stoichiometry (1) and thermodynamics (2)	Enzyme identification tool	Toxicity of intermediates	Host or organism specificity	Availability	References
BNICE*	Yes	GCM ¹⁵	Yes (1,2)	BridgIT ¹⁶		Yes	Open data via ATLAS ¹⁷	Tool development ⁴ , applications ^{5,18-23} , reviews ^{10,24,25} , experimental validation ²⁶
RetroPath series**	Yes	Yes	Yes (1)	Selenzyme ^{27,28}	Yes ²⁹	Yes	Open-source & open data via XTMS	Tool development ^{6,7,30,31} , reviews ³² , pathway search ³³ , experimental validation ³⁴
novoStoic	Yes	Yes ³⁵	Yes (1,2)	Yes		Yes	Open-source	Tool development ⁸ , reviews ^{9,13}
GEM-Path	Yes	GCM ¹⁵	Yes (1,2)	Yes		Yes		Tool development ³⁶
SimPheny	Yes	Yes ³⁷		Manual		Yes	Commercial	Tool development (Genomatica) ³⁸ , experimental validation ³⁹
ReactPRED	Yes	Yes					Open-source	Tool development ⁴⁰
Transform-MinER	Yes			Yes			Online tool	Tool development ⁴¹
DESHARKY				Yes		Yes		Tool development ⁴² , review ⁴³
ReBiT	Yes	Yes						Review ⁴⁴
Cho et al.	Yes	Yes						Tool development and application ⁴⁵

*The development of BNICE has started at Northwestern University in 2004, and later been continued at EPFL under the new name BNICE.ch. Here, publications based on both versions are included.

**Comparison includes the original RetroPath method as well as its extended versions RetroPath2.0 and RetroPath RL

5.1.1 Iterative network generation

Retrobiosynthesis starts with defining the biochemical search space: What kind of metabolites and reactions should be considered? In the most restricted setting, the search is limited to known metabolites and reaction of a specific organism. From there, the search space can be expanded to all known biological compounds and reactions, then to novel reactions, to compounds only found in chemical databases, and finally it can even include novel molecular structures. Depending on the question we try to address with retrobiosynthesis, and on how much biochemical knowledge is available for the target compound, the search space will be delimited differently. In the standard case, the target is a known biological or chemical compound, and we will consider all known chemical and biological compounds, as well as all known and predicted biochemical reactions. Known compounds and reactions can be collected from databases such as KEGG⁴⁶, MetaCyc⁴⁷ and BRENDA⁴⁸, and chemical compounds can be taken from chemical databases such as PubChem⁴⁹. In the following, we will refer to compounds that are only present in chemical databases as “chemical compounds”. However, not all molecules present in nature can be found in biological database, and they may only exist in chemical databases, meaning that a “chemical compound” in our definition may still be of biological origin. Finally, we need to define the type of biochemistry considered in the network generation by choosing an appropriate set of generalized, biochemical reaction rules. While for specific questions it can be advantageous to restrain the possible biochemistry to certain types of EC classes or to enzymatic rules already known to be present in the host organisms, we generally opt for the whole set of available reaction rules to cover all of known biochemistry.

Once the search space is defined, we apply a network generation algorithm to expand a hypothetical reaction network around the target compound (Figure 5.2). In a first iteration, the reaction rules are applied one by one to the target compound, thus generating all possible reactions and products that lie within the search scope. The products of the first iteration are then analyzed to determine whether or not they should be subject to further network expansion. For example, we can decide to only further follow compounds with a specific elemental composition or a given maximal molecular weight. We can also look at the ratio of conserved atoms between the target and the product, and only validate reactions conserving a minimal ratio of atoms. These metrics can be used to designate products that are eligible for further network expansion. These eligible products are then used as substrates in a second iteration of reaction generation, and so on, until a predefined number of iterations is reached, or until the computational resources are exhausted. In the standard case, the resulting network contains known and novel, hypothetical reactions between known biological and chemical molecules.

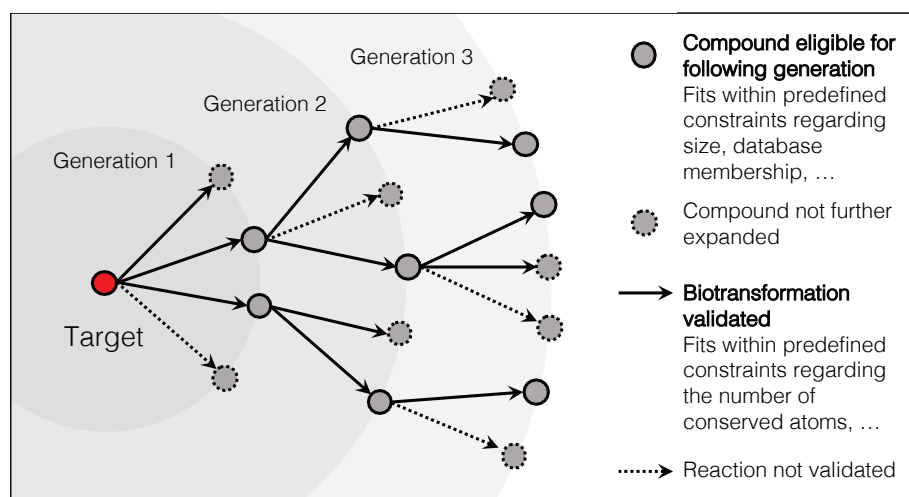


Figure 5.2: The BNICE.ch network generation process.

Alternatives to BNICE.ch for network generation with continued developments are the previously mentioned RetroPath^{6,7} and novoStoic^{8,9}. RetroPath relies on a set of automatically generated reaction rules, where the atomic diameter of the rule is a parameter that needs to be defined *a priori*. NovoStoic extracts its reaction rules based on an algorithm termed rePrime that encodes reaction centers using a prime factorization-based technique. Another related tool, called enviPath has been specifically designed for biodegradation pathways and it employs a set of expert-curated biodegradation rules^{50,51}. While all of these tools rely on the concept of generalized reaction rules, their network generation processes differ in terms of available parameters and user interaction with the algorithm.

5.1.2 Pathway search within hypothetical metabolic network

Once we have generated a hypothetical biochemical network, we want to extract possible sequences of reactions that connect the target compound to back to precursor compounds in the host organism. The extraction of pathways can be achieved by solving an optimization problem, or by finding paths in a graph-based representation of the metabolic network, as reviewed by Wang *et al.*¹³. Here, we employ an atom-conserving, graph-based pathway search termed NICEpath and explained in detail in Chapter 3.

NICEpath takes as input the network generated by BNICE.ch, where each reaction is annotated with a metric that defines substrate-product pairs, weighted based on the number of atoms conserved between the substrate and the product. The Conserved Atom Ratio (CAR) is used to construct a searchable, weighted graph, where compounds are nodes, biotransformations are edges, and where edge weights represent the conservation of atoms in a given biotransformation. The weights are inversed values of the CAR, meaning that substrate-product pairs with a high number of conserved atoms close to each other in the graph, and pairs that only conserve few atoms are far away from each other. This

representation enables us to apply a powerful graph search method to extract the shortest pathways from the overall reaction network, biased towards atom-conserving routes.

The following parameters can be set by the user to fine-tune the pathway search algorithm: (i) Maximal number of reaction steps and maximal number of pathways, (ii) the maximum number of alternative reactions to be returned for each edge (*e.g.*, alternative cofactor usages for a given biotransformation), (iii) the model identifier of the organism to which we want to connect the input compound(s). The resulting pathways are automatically ranked based on the sum of the individual distances in each reaction. NICEpath further provides information on the average CAR in the pathway, the number of known, novel and total reaction steps involved. For a more detailed discussion of NICEpath and other pathway search tools, the reader is referred to Chapter 3 of this thesis.

5.1.3 Stoichiometric and thermodynamic pathway feasibility

In a next step, we evaluate the stoichiometric and thermodynamic feasibility of each pathway in the GEM of the chassis organisms. This procedure is necessary to remove infeasible routes and to rank the remaining routes based on yield. A metabolic model of the host organism is a pre-requisite for the following analyses.

A first criteria of feasibility is that the compounds participating in the pathway are balanced with regard to the organisms' metabolism in terms of cofactor usage and co-substrate availability. To assess the stoichiometric feasibility, each predicted pathway is appended to the GEM of the chassis organism, and the production of the target compound is optimized in a Flux Balance Analysis⁵² (FBA) type of problem. If the target compound can be produced, the pathway is stoichiometrically feasible, and the maximum theoretical yield from the main carbon substrate is calculated.

A second criteria is the thermodynamic feasibility of the production of the target compound. In order to evaluate the thermodynamic feasibility, thermodynamic data has to be collected or calculated for the compounds and reactions in the pathway. A common approach to estimate the thermodynamic properties is the Group Contribution Method (GCM)¹⁵. The estimations of the Gibbs free energy of formation and the Gibbs free energy of reaction are a prerequisite to perform Thermodynamics-based Flux Analysis^{53,54} (TFA) on the production of the target molecule in the chassis model. We determine whether the compound can be produced from a thermodynamic point of view by optimizing the production of the target in TFA, and we calculate the maximum theoretical yield in the thermodynamically constrained model.

5.1.4 Finding enzymes for predicted reactions

Once the pathways are established and their feasibility evaluated in the GEM of the host organism, we need to ensure that each step in the pathway can be catalyzed by an enzyme. For each reaction, three cases are possible: (i) The reaction is known, and biological databases provide information on the enzyme catalyzing the reaction, (ii) the reaction is known,

but no enzyme has been associated yet to the biotransformation, or (iii) the reaction is novel, not known to any database, and predicted by a BNICE.ch reaction rule. Reactions belonging to the latter two cases are generally called “orphan”, and we need to rely on chem- and bioinformatic tools to find a known enzyme that is either promiscuous enough to catalyze the novel reaction, or that could be engineered to perform the desired biotransformation. These tools can also be applied to known reactions with associated enzymes (first case) if one is looking for alternative enzymes that are more compatible with the host organism, or if one is seeking to explore the promiscuous reaction space of an enzyme.

Enzyme prediction relies on the comparison of the molecular structures of the reaction and its participating reactants by using so-called “reaction fingerprints”, which represent the reaction in a condensed string of characters. These fingerprints are then compared among each other, and similarity measures such as the Tanimoto distance can be employed to score the structural similarity between the reactions⁵⁵. In practice, one first establishes a reference database listing all known reactions, their associated enzymes and their reaction fingerprints. Next, the fingerprint is calculated for each novel query reaction and compared to the reference database. Reactions showing high structural similarity are retrieved along with their similarity score and their associated enzymes.

In our workflow, we use the computational tool BridgIT to associate known enzymes to the novel reactions generated by BNICE.ch¹⁶. BNICE.ch already provides information on the type of reaction by defining the EC number up to the third level. The information on the reactive site stored in the reaction rule is further used to center the fingerprint generation on the reactive site and its surrounding structure. These reactive-site centric fingerprints are then used to find the known reactions that are most similar to the novel query reactions in terms of reactivity, reactive site, and molecular structure. The known reactions and associated EC numbers can be easily mapped to protein and gene sequences using standard bioinformatic resources such as UniProt⁵⁶.

Alternative methods such as EC-BLAST⁵⁷, Selenzyme²⁸ or E-zyme⁵⁸ employ similar cheminformatic approaches to identify enzymes for orphan reactions. Like BridgIT, EC-BLAST and E-zyme rely on reactive-site centric comparison, which has been shown to be crucial to correctly identify promiscuous enzymatic activity.

5.1.5 Ranking, visualization and availability

Each step in the retrobiosynthesis workflow adds its own set of scores to the resulting pathways: The network generation provides information on the nature of the participating compounds and reactions in the network, *i.e.* the size of the compounds, whether the compounds and reactions can be found in any biological or chemical databases, the type of biochemistry involved, etc. The pathway search outputs the length of the pathway as well as a metric for atom conservation throughout the pathways. The stoichiometric and thermodynamic analyses calculate the theoretical maximal yield and tell us whether or not a given pathway is feasible within the host organism. Finally, enzyme annotation tools calculate a

similarity score between novel, predicted reactions and known enzyme-catalyzed reactions to estimate the feasibility of biocatalysis for the predicted reactions. Given all of these scores and rankings, it is up to the scientist to decide which ones should be prioritized in the final ranking of pathways, depending on the specific case of application and the objective of the project. For this reason, the final ranking should be individually adapted for each retrobiosynthesis project, taking into account the specific needs and constraints of the metabolic engineering task.

A recurrent criticism of retrobiosynthesis tools is that compared to the number of available tools, only a small fraction of the predictions has been validated experimentally. Potential reasons for this are (i) the difficult accessibility of tools, which are either closed-source or not easy to use by experimentalists without background in informatics, and (ii) the lag time between prediction and experimental validation. For example, the first hypothetical reactions predicted in the ATLAS of Biochemistry, published in 2016, have been validated in an experimental study only recently²⁶. This example shows that it is crucial to invest in the accessibility of tools and, even more so, in data visualization for communicating effectively with experimental collaborators.

While other retrobiosynthesis tools opt for full open-source solutions, the BNICE.ch framework is closely tied to an infrastructure of communicating databases on an in-house server. Furthermore, our overall retrobiosynthesis framework involves the application of different tools such as BridgIT and pathway evaluation within the host model, which require a trained expert to optimize each step in the workflow towards the bioengineering goal (*e.g.*, choice of chassis, consideration or exclusion of specific reaction mechanisms, compounds to be avoided due to patenting conflicts). Hence, we believe that providing curated retrobiosynthesis results online in a user-friendly, open-access manner is currently the best solution to share our data. All of our predicted and published pathways can be consulted and manually inspected under <https://lcsb-databases.epfl.ch/pathways/GraphList>.

5.2 Retrobiosynthesis for 1,4-butanediol and bisabolene

The results of this Subchapter were generated with the help of Omid Oftadeh, who performed the stoichiometric and thermodynamic analysis of the pathways in, and Homa Mohammadi-Peyhani, who predicted enzymes with BridgIT. The yeast model used for pathway evaluation has been thermodynamically curated by Maxime Curvat.

The following Subchapter is dedicated to the exemplar illustration of a typical retrobiosynthesis project, with the aim of benchmarking our retrobiosynthesis workflow. As a showcase, we chose the two industrially important compounds 1,4-butanediol (1,4-BDO) and bisabolene. Both compounds have been the subject of several successful metabolic engineering efforts, yet their chemical structures are very different. Here, we apply the state-of-the-art retrobiosynthesis workflow to both of these compounds, and we compare our results with information available in biochemical databases and in scientific literature.

1,4-BDO is a major commodity chemical that is widely used in industry for the chemical synthesis of different plastics, elastic fibers and pharmaceuticals, with a global market of over 2.5 million tons per year. It is not native to any known organism, and traditionally sourced from petroleum-based feedstocks. The bioproduction of 1,4-BDO has been achieved for the first time by Yim *et al.* in 2011, using *Escherichia coli* (*E. coli*) as a host organism³⁹. In 2015, Liu and Lu reported a more advanced metabolically engineered system for autonomous 1,4-BDO bioproduction in *E. coli* from *D*-xylose that integrated genetic control tools from synthetic biology⁵⁹. Another biosynthetic route was introduced by Wang *et al.* in 2017 that makes use of a rationally redesigned diol dehydratase⁶⁰.

The chemical bisabolene is a natural product belonging to the chemical group of terpenes, also called isoprenoids⁶¹, and its chemical properties make it a good candidate for biosynthetic alternative to diesel fuel⁶². Bisabolene is naturally produced by some groups of fungi⁶³, and it has also been found in essential oils extracted from lemon and from different member of the lamiaceae family, where it contributes to the balsamic odor. The natural biosynthesis of bisabolene has first been characterized in *Abies grandis*, where the compound is produced from farnesyl diphosphate by the enzyme bisabolene synthase⁶⁴. Farnesyl diphosphate is produced from the precursors isopentenyl pyrophosphate (IPP) and dimethylallyl pyrophosphate (DMAPP), which are synthesized either via the mevalonate pathway, or via the non-mevalonate pathway, also called the MEP pathway. A synthetic route for the biosynthesis of isoprenoids has recently been discovered by Clomburg *et al.* and termed isoprenoid alcohol pathway⁶⁵.

Bisabolene has been the target of several bioengineering studies in the past. In 2011, *E. coli* and *Saccharomyces cerevisiae* (*S. cerevisiae*) have been engineered through to produce bisabolene from simple sugars for the first time⁶². In contrast to the engineered biosynthesis of 1,4-BDO involving the design of novel biochemical reactions, the bisabolene bioproduction in yeast was achieved through heterologous expression of the bisabolene synthase from *Abies grandis*. Biosynthesis of the isoprenoid has further been reported in a the

cyanobacterium *Synechococcus* sp. PCC 7002 and in *Streptomyces venezuelae*^{66,67}. Bisabolene exists in the form of the different isomers α -, β - and γ -bisabolene, which differ in the position of the double bond close to the ring. Since the engineered strains all produce α -bisabolene, we focused our study on this specific isomer. On top of that, there are several stereoisomers possible due to two stereocenters in the molecule, but which are not further distinguished for the computational analysis.

In the following, we perform an overall, state-of-the-art retrobiosynthetic analysis of these two compounds, starting with the exploration of the biochemical space around the target molecules.

5.2.1 Generation of biochemical reaction network around target compounds

To generate a hypothetical reaction network around the target compounds, we applied the BNICE.ch retrobiosynthetic network generation on the two target compounds 1,4-BDO and bisabolene. The following constraints were applied at this stage: The network generation was limited to seven iterations, and only compounds known to the LCSB database were allowed to be produced. Also, compounds eligible for the next generation needed to (i) have maximum two carbon atoms more than the target compound, and (ii) be connected to an eligible compound from the previous generation via a biotransformation with a CAR higher than or equal to 0.34, a cutoff previously found to be optimal to predict biologically relevant substrate-product pairs (see Chapter 3). Finally, the algorithm would not start a new generation when the general limit of 100,000 reactions or 20'000 compounds is reached to avoid using up all of the allocated memory resources in the following generation.

Given these constraints, BNICE.ch predicted a reaction network of 22,765 compounds and 122,139 reactions for 1,4-BDO, and a network of 48,852 compounds and 207,634 reactions for bisabolene (Figure 5.3). The results show that the network generation evolved differently for the two compounds. The main reason for this difference is the upper limit on the number of carbons, which was set to six for 1,4-BDO and to 17 for bisabolene. Hence, the latter produced a bigger network due to a higher number of possible chemical structures. As a result, the network generation algorithm stopped after six generation for bisabolene because the memory resources were used up.

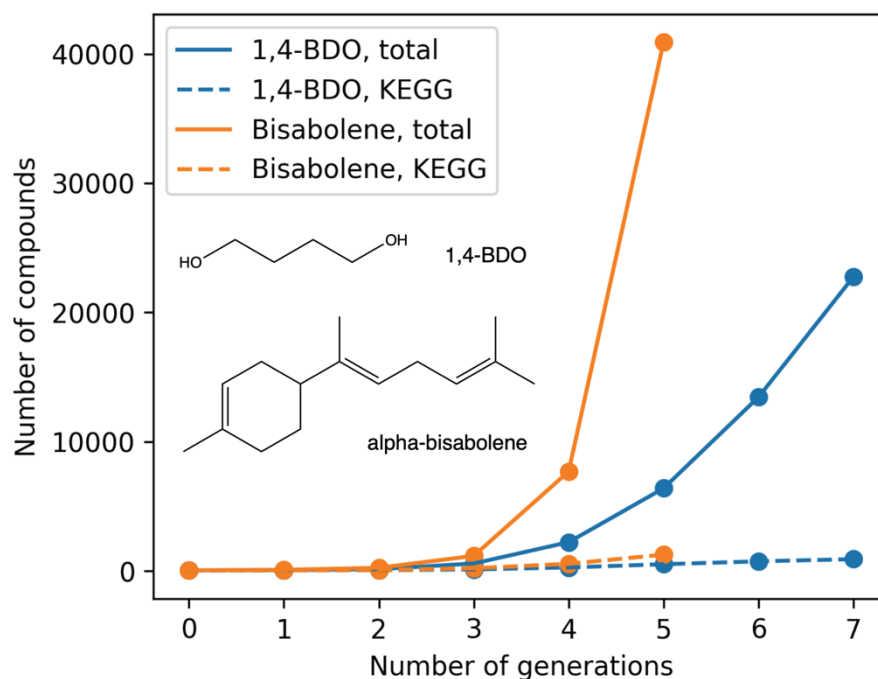


Figure 5.3: Compounds produced in each BNICE.ch iteration during the network generation process for 1,4-BDO (blue) and bisabolene (orange). Solid lines indicate the total number of compounds, and the dashed lines show the number of biological KEGG compounds.

5.2.2 Finding pathways to host precursors

Once we had a network of known and novel, hypothetical reactions around the target compounds, we tried to connect each target to all precursors in common between the generated network and the genome-scale model of the host organisms. For 1,4-BDO, the only host organisms engineered for bioproduction was *E. coli*, represented in our approach by the genome-scale model iJO1366⁶⁸. Bisabolene bioproduction has been engineered in four organisms in the past; *E. coli*, *S. cerevisiae*, *Synechococcus* sp. PCC 7002 and in *Streptomyces venezuelae*. The first three organisms are here represented by iJO1366, iMM904⁶⁹ and iJB785 (for *Synechococcus elongatus* PCC 7942)⁷⁰, respectively. Unfortunately, no genome-scale model was available for *Streptomyces venezuelae*.

For each precursor found in the respective models, we analyzed the shortest 100 paths connecting the precursor to the target, considering a maximum of six alternative reactions per edge in the network. From the resulting pathways, we ignored those that required a co-substrate not present in the organism. We further removed pathways that produced molecular oxygen and that fixed carbon dioxide, since these reactions are generally strongly unfavorable in terms of thermodynamics, with estimated Gibbs free energies of reaction of generally more than ~10 kcal/mol. Finally, no native *E. coli* intermediates were allowed in the pathway to reduce redundancy between the pathways.

For 1,4-BDO, we found a total of 6,313 pathways for 155 precursor compounds from *E. coli*. The shortest pathway connects 1,4-BDO to butanol in two reaction steps, and the longest

connects the target to 5-Phosphoribosyl diphosphate in nine reaction steps. For bisabolene, we found 4,468 pathways for 71 precursor compounds in *E. coli*, with the number of reaction steps ranging from one to eleven. We further found 4,187 pathways for 83 precursor compounds in *S. cerevisiae*, and 4,474 pathways for 72 precursor compounds in *Synechococcus*, with the same range of pathway length as for *E. coli*. The precursor compounds found in the different species were largely the same: 67 precursors were found in all the GEMs of the three species. 16 precursors were unique to yeast, while four precursors were only found in *E. coli* and *Synechococcus*, and one precursor was unique to *Synechococcus*. The set of pathways found for *Synechococcus* and *E. coli* were exactly the same, with one single exception; *Synechococcus* had 7 extra pathways for the precursor D-1-Aminopropan-2-ol O-phosphate. Since the results for *E. coli* and *Synechococcus* were almost identical, and given the fact that there was no in-house, thermodynamically curated GEM for *Synechococcus*, we decided to perform the stoichiometric and thermodynamic pathway evolution only within *E. coli* and yeast GEMs. In general, the pathways for 1,4-BDO tended to be shorter, which is due to the fact that the structure of the molecule is simpler and closer to the central carbon metabolism (Figure 5.4). The pathway length distribution for bisabolene turned out to be very similar between *E. coli* and yeast. We had a closer look at these pathways and we found that 3,280 pathways (78%) were identical between *E. coli* and yeast. Comparing these pathway properties can give us a first impression on which organisms might be more suited than others for the bioproduction of a given compounds. However, a stoichiometric and thermodynamic analysis is needed to ensure the feasibility of the pathways within the proposed host organisms.

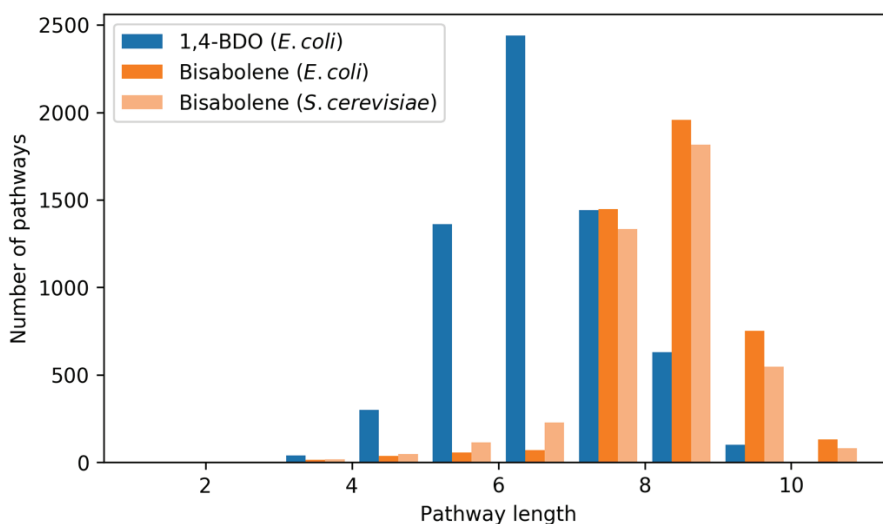


Figure 5.4: Distribution of pathway lengths for 1,4-BDO in *E. coli* and for bisabolene in *E. coli* and *S. cerevisiae*.

5.2.3 Stoichiometric, thermodynamic and biocatalytic pathway evaluation

To determine the stoichiometric feasibility of the pathways, we inserted them one by one into the GEM models of the respective organisms, and we performed FBA to see if the compound can be produced. In *E. coli*, 3,695 out of 6,313 pathways for 1,4-BDO were stoichiometrically feasible, starting from 102 out of the initial 155 precursors. For bisabolene, 3,374 out of 4,187 pathways were FBA feasible in *E. coli*, for 52 out of 71 precursors. In yeast, 2,578 out of 4187 pathways to bisabolene were feasible, for 52 out of 83 precursors. We further reduced the number of pathways by filtering out those that were not feasible from a thermodynamic point of view. For the thermodynamically feasible pathways we also calculated the maximum possible yield per glucose molecule. The TFA analysis left us with 325 pathways towards 17 different precursors for 1,4-BDO in *E. coli*, 1,417 pathways towards 32 precursors for bisabolene in *E. coli* and 52 pathways towards 10 precursors in yeast (Table 5.2). Given that the original sets of pathways for bisabolene are very similar for *E. coli* and yeast, we will need further investigation into the thermodynamic properties of the two models to explain the remarkable difference in the number of feasible pathways.

A closer look at the shortest pathways for each target revealed differences in how the compound was connected to the host metabolism (Figure 5.5): 1,4-BDO connected to the chemically related native *E. coli* metabolite 4-hydroxybutanoic acid in only three steps, and to 3-hydroxypentanoate and 3-methyl-2-oxopentanoate in four steps. Bisabolene connected to trans, trans-farnesyl diphosphate in one, two or three reaction steps, and additionally to dimethylallyl diphosphate within four reaction steps. Interestingly, these two compounds are known biosynthesis products of the MEP pathway which is active in *E. coli* and produces the precursors for the biosynthesis terpenoid backbone. In yeast, terpenoids are produced via the mevalonate pathway. Accordingly, the shortest pathways found in yeast connect to the mevalonate pathway intermediates (*R*)-5-phosphomevalonate and (*R*)-mevalonate.

In a previous retrobiosynthesis study on 5-methylethylketone (MEK), 3,679,610 novel pathways towards five precursors were reconstructed from a biochemical reaction network generated by BNICE.ch⁵. Out of these novel pathways, only 487,411, or 13.2%, were found to be stoichiometrically feasible and 18,622, or 0.5%, were thermodynamically feasible. In contrast, 5.1% and 33.9% of pathways reconstructed by NICEpath were found to be feasible for 1,4-BDO and bisabolene, respectively. The difference between these two studies is that the previous approach performed a pathway search algorithm on an unweighted, undirected graph that did not take into account atom conservation between substrate-product pairs. Even though the results are not directly comparable, this outcome still confirms the value of the atom-conserving graph search approach employed in NICEpath.

Table 5.2: Stoichiometric and thermodynamic filtering of pathways found by NICEpath in *E. coli*. Abbreviations: # PWs - Number of pathways, % PWs - Percentage of pathways kept after filtering with respect to the total number of pathways, # prec - Number of precursors)

	1,4-BDO – <i>E. coli</i>			Bisabolene - <i>E. coli</i>			Bisabolene - Yeast		
	# PWs	% PWs	# prec	# PWs	% PWs	# prec.	# PWs	% PWs	# prec
All pathways	6,313	100.0%	155	4,183	100.0%	71	4,187	100.0%	83
FBA feasible	3,695	58.5%	102	3,774	90.2%	52	2,578	61.6%	52
TFA feasible	325	5.1%	17	1,417	33.9%	32	52	1.2%	10

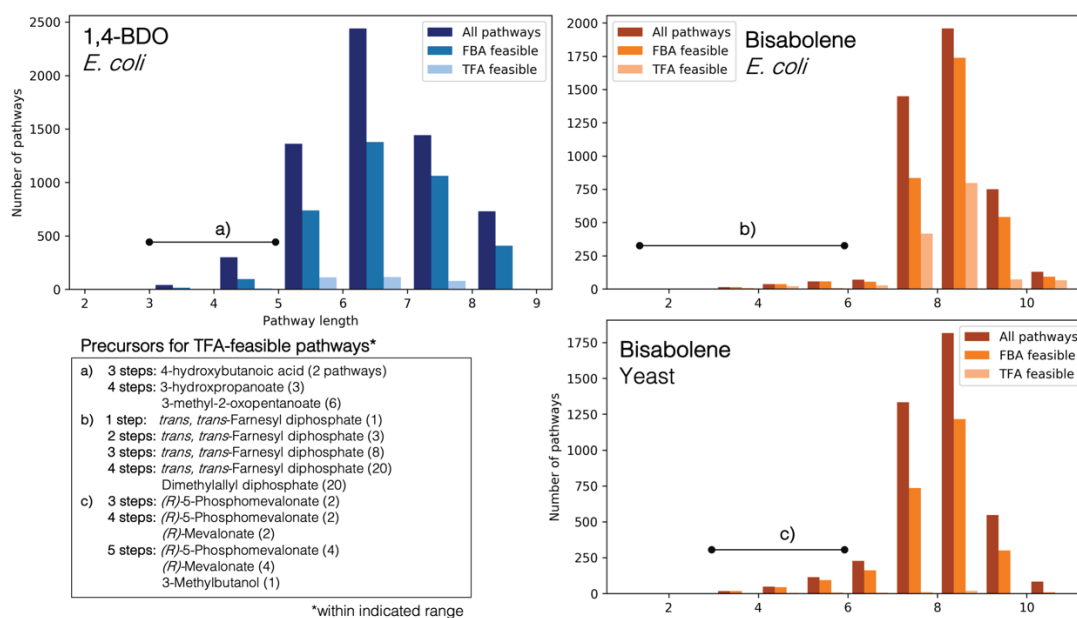


Figure 5.5: Distribution of pathway length after stoichiometric and thermodynamic curation, respectively. Precursor compounds of the shortest pathways are listed for each analysis.

Finally, each reaction step in a pathway needs to be catalyzed by enzyme. For known reaction steps, suitable enzymes can be found in biochemical databases, but for novel reaction steps, an enzyme prediction tool like BridgIT is indispensable. Here, we applied BridgIT on all the reactions to find all the known enzymes that may have the biochemical potential to catalyze the desired steps. In case the reaction is already known, alternative enzymes can still be interesting in case the known enzyme does not express well in the host, or if the known enzyme creates patent-related issues. We extracted the top five hits BridgIT hits for each reaction, not considering hits with a BridgIT score lower than the recommended threshold of 0.3, below which BridgIT results have been shown to not be significant anymore¹⁶. For each pathway, we calculated the average BridgIT score by considering the highest score for each reaction step in order to provide an overall metric assessing the availability of enzymes.

5.2.4 Pathway ranking, visualization and comparison

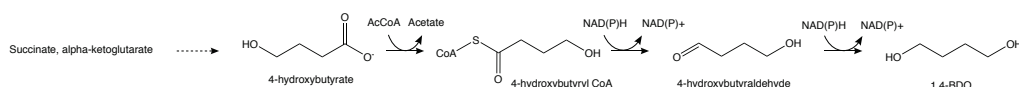
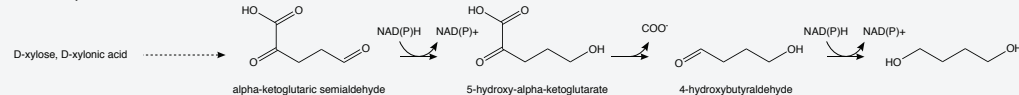
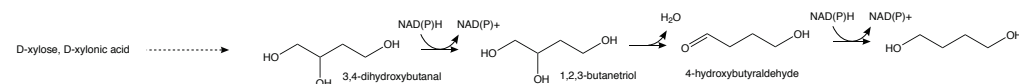
The result of our retrobiosynthesis approach is a list of theoretically feasible pathways leading from precursors present in the host organism to the target compound. Pathways are annotated with a set of metrics consisting of a pathway length score (the inverse of the number of reaction steps), the average atom conservation, the maximum yield from glucose, and a score for enzyme availability. An overall ranking is achieved by adding the different scores, each one ranging from zero to one. The ranked pathways have been uploaded to our website (<https://lcsb-databases.epfl.ch/pathways/GraphList>), where users can consult the pathways and re-rank them based on their own criteria.

First, we examined the 325 thermodynamically feasible pathways for 1,4-BDO, and we selected top-ranking pathways to illustrate the influence of the different ranking criteria on the overall pathway score (Table 5.3). However, when we compared the engineered biosynthetic pathways reported in literature to our pathways, none of the reported pathways could be found in our results (Figure 5.6). Possible reasons for this might be that the necessary reaction mechanisms were not present in the set of applied BNICE.ch reaction rules, that the pathway search could not find these pathways or that FBA or TFA considered these pathways as infeasible. Additional investigations will be required to settle this question. Nevertheless, our workflow could find interesting alternative routes for the production of 1,4-BDO.

Table 5.3: Overview on top pathways for the bioproduction of 1,4-BDO in *E. coli*, ranked by their overall score. The yield is calculated for glucose.

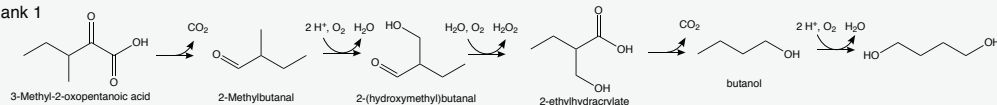
Rank	Precursor	Pathway length	1 / Pathway length	Known reaction score*	Mean Bridgl T score	Max yield [mol/mol]	Max yield [g/g]	Average CAR	Overall score
1	(S)-3-Methyl-2-oxopentanoic acid	5	0.2	0.2	0.670	0.435	0.218	0.69	2.195
2	(S)-3-Methyl-2-oxopentanoic acid	5	0.2	0.2	0.669	0.435	0.218	0.69	2.194
3	(S)-3-Methyl-2-oxopentanoic acid	5	0.2	0.2	0.637	0.435	0.218	0.72	2.192
...	(S)-3-Methyl-2-oxopentanoic acid								
11	(S)-Malate	5	0.2	0	0.632	0.535	0.268	0.73	2.097
12	(S)-Malate	5	0.2	0	0.642	0.535	0.268	0.71	2.087
13	(S)-Malate	5	0.2	0	0.616	0.535	0.268	0.73	2.081
14	4-Hydroxybutanoic acid	3	0.333	0	0.733	0.448	0.244	0.56	2.074
15	L-Aspartate	6	0.167	0	0.610	0.527	0.264	0.76	2.064

*Number of known reactions / pathway length

Yim *et al.*, 2011Liu *et al.*, 2015Wang *et al.*, 2017

Top ranked pathways

Rank 1



Rank 14

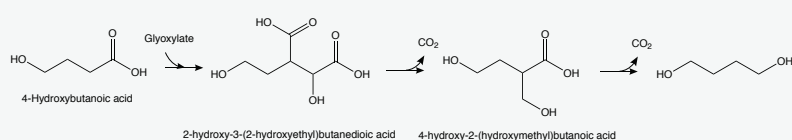


Figure 5.6: Comparison of the three synthetic 1,4-BDO bioproduction pathways in *E. coli* with two of the top-ranked pathways in this study. Protonation states and corresponding names of compounds originating from BNICE.ch are not corrected for pH 7. For example, 4-hydroxybutanoic acid corresponds to 4-hydroxybutyrate in standard biological conditions.

Second, we examined the top-ranked pathways for bisabolene in *E. coli* (Table 5.4) and in yeast (Table 5.5). Since both organisms are capable of producing the immediate precursor of bisabolene, farnesyl diphosphate (FPP), it is not surprising that NICEpath finds the shortest pathways from this metabolite in *E. coli* (Figure 5.7). In yeast, the top-ranking pathways start from 5-phosphomevalonate and mevalonate, both intermediates from the mevalonate pathway active in yeast, and produce bisabolene from farnesyl phosphate. This is surprising, since FPP is known to be present in yeast. A closer investigation revealed that the pathway from FPP to bisabolene has been found by NICEpath, but filtered out because the pathway was not thermodynamically feasible in the yeast GEM. A more detailed analysis of the thermodynamic properties of the yeast model will hopefully answer why the production from FPP was not considered feasible.

Table 5.4: Top five pathways for bisabolene in *E. coli*. The yield is calculated for glucose.

Rank	Precursor	Pathway length	1 / Pathway length	Known reaction score*	Mean BridgIT score	Max yield [mol/ mol]	Max yield [g/g]	Average CAR	Overall score
1	FPP	1	1	0	1	0.285	0.323	0.51	2.795
2	FPP	2	0.5	0.5	0.734	0.285	0.323	0.76	2.779
3	FPP	3	0.333	0.333	0.954	0.285	0.323	0.65	2.556
4	FPP	3	0.333	0.333	0.954	0.283	0.321	0.65	2.554
5	FPP	3	0.333	0.333	0.871	0.285	0.323	0.63	2.453

*Number of known reactions / pathway length. FPP: Farnesyl diphosphate

Table 5.5: Top five pathways for bisabolene in yeast. The yield is calculated for glucose.

Rank	Precursor	Pathway length	1 / Pathway length	Known reaction score*	Mean BridgIT score	Max yield [mol/mol]	Max yield [g/g]	Average CAR	Overall score
1	5-PMev	3	0.333	0	0.930	0.250	0.284	0.55	2.063
2	5-PMev	4	0.25	0	0.812	0.250	0.284	0.66	1.972
3	5-PMev	5	0.2	0	0.889	0.250	0.284	0.57	1.909
4	(<i>R</i>)-Mevalonate	4	0.25	0	0.899	0.250	0.284	0.50	1.899
5	5-PMev	5	0.2	0	0.882	0.246	0.279	0.57	1.898

*Number of known reactions / pathway length. 5-PMev: (*R*)-5-Phosphomevalonate

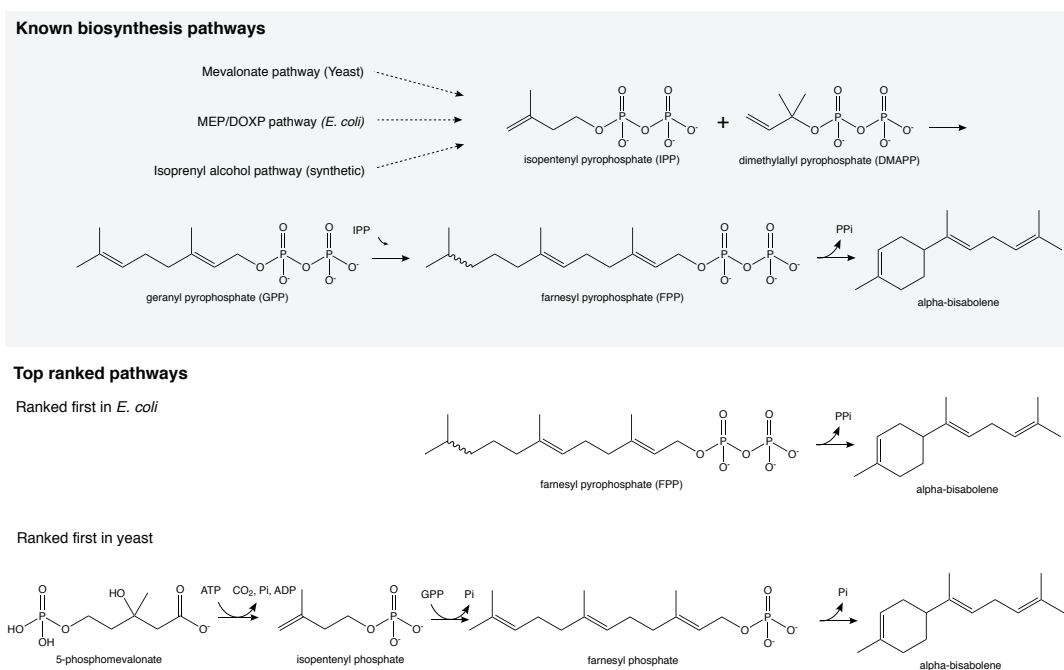


Figure 5.7: Biosynthesis pathways for bisabolene in nature, and top-ranked pathways in *E. coli* and yeast. Abbreviations: Inorganic phosphate (Pi) and diphosphate (PPi).

5.2.5 Conclusion

This study illustrates a state-of-the-art retrobiosynthetic approach for the discovery of new biosynthesis pathways for the production of chemicals of industrial interest. It should be mentioned at this point that further analyses can be applied to narrow down the space of feasible pathways, for example by minimizing the loss of carbon to CO₂, by including kinetic constraints, or by biasing the ranking towards a specific engineering goal, for example to circumvent patents by excluding affected enzymes and pathway intermediates. Furthermore, it should be emphasized that this approach is unbiased regarding the type of biochemistry considered and the selection of precursor compounds. The only human intervention in our retrobiosynthesis workflow is, for the moment, the choice of the chassis organism. Future developments of the workflow will address this matter by proposing an automated screen that is able to select the best-suited host from a range of organisms.

5.3 Exploring the chemodiversity around the noscapine pathway

This subchapter is the result of a collaboration with the experimental lab of Professor Christina Smolke at the University of Stanford, which started in the context of a four-months exchange sponsored by the private company Firmenich. The subchapter is a condensed version of manuscript, to be submitted. Experimental results have been obtained by Dr. James Payne, and BridgIT results have been provided by Homa Mohammadi-Pehani.

Plants synthesize a remarkable range of complex and valuable molecules, known as plant natural products (PNPs), commonly used as flavors, fragrances, and medicines. However, production of these molecules via extraction from plant biomass can be 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 significantly higher fraction are derivatives of natural products⁷¹. The range of PNP derivatives accessible to researchers is typically limited to those that can be easily produced chemically from PNPs extracted from plants, while we can envision many more potential derivatives that could be made via regioselective enzymatic functionalization 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 artemisinic acid, a precursor to the antimalarial artemisinin, in *Saccharomyces cerevisiae* in 2006⁷², there has been a rapid increase in the size and complexity of pathways expressed heterologously.¹⁴ This has been demonstrated clearly by the progress made on the bioproduction of benzylisoquinoline alkaloids (BIAs), a class of PNPs of particular medicinal interest, with members providing analgesic, antitussive, and anticancer effects. In 2015, the *de novo* biosynthesis of the BIAs thebaine and hydrocodone in *S. cerevisiae* was reported⁷³, the latter of which is 11 enzymatic steps from endogenous metabolites, and in 2018 the *de novo* biosynthesis was reported for the non-opioid BIA noscapine⁷⁴, 16 enzymatic steps from endogenous metabolites. Halogenated derivatives of tyrosine were fed to the engineered yeast strains comprising the reconstructed noscapine biosynthetic pathway to produce halogenated derivatives of noscapine intermediates. However, the non-native substrates were not tolerated as well as the native substrates of the pathway enzymes, and as such derivatives of only early intermediates in the pathway were produced. Accessing derivatives of later pathway intermediates with a feeding strategy will thus require engineering of key pathway enzymes to accommodate the new substrates or feeding of derivatives of later intermediates, which are of increasing chemical complexity and often not commercially available.

An alternative approach to produce derivatives of PNPs and their intermediates would be to integrate additional enzymes into microbes 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 for other than their natural function, and in most cases an enzyme will not be known *a priori* with the desired non-native function. 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.

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 5.8). In contrast to previously reported retro-biosynthesis 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 biosynthetic chemistries. As the chemical space is large for even a pathway of modest complexity, we implemented a filtering strategy by ranking the full set of potential compounds by the number of PubMed and patent citations reported to prioritize compounds that have been previously explored for their biomedical potential. Other ranking algorithms can readily be employed in the workflow, and compounds that have not appeared in previous reports could be prioritized for their potential novelty. The method then identifies enzymes capable of carrying out the desired transformations on the prioritized set of compounds using the enzyme prediction tool BridgIT. Finally, the top predicted enzyme candidates are experimentally characterized to validate those capable of producing the target molecule.

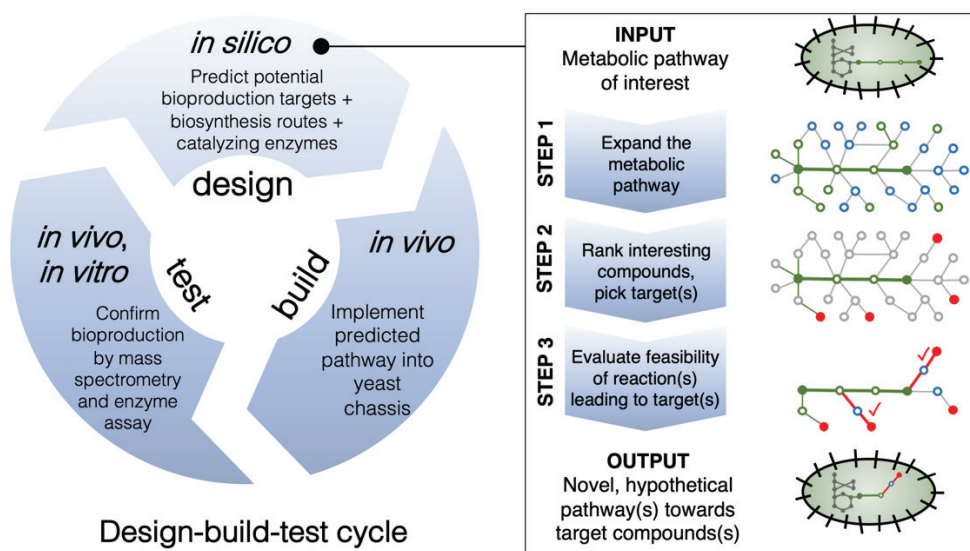


Figure 5.8: Overall workflow. Left: Applied design-build-test cycle. Right: Computational workflow. Circles represent compounds, edges represent biotransformations. Green is used to designate known biological reactions and compounds, blue circles are compounds from the chemical space without specific biological annotation, and red circles show compounds selected for their popularity in scientific literature and in the patent landscape.

We applied the described workflow to the reconstructed noscapine biosynthetic pathway in yeast. Using the workflow, we narrowed our search to enzyme candidates capable of producing (*S*)-tetrahydropalmatine, a PNP found in plants of the genus *Corydalis* widely used in Chinese herbal medicine and not previously produced from heterologous BIA pathways. (*S*)-Tetrahydropalmatine has been shown to possess analgesic and anxiolytic effects and has shown promise as a potential treatment for opiate addiction^{75–77}. After experimental evaluation of the top six 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. In addition, one of the screened enzymes was shown to produce an N-methylated derivative of (*S*)-tetrahydrocolumbamine. To our knowledge, our work describes the first use of a computational workflow to produce a novel product from a heterologous biosynthetic pathway. As the number of reconstructed heterologous pathways for PNPs continues to increase, we anticipate that the described workflow can be used to produce many novel, chemically complex compounds spanning diverse therapeutic activities.

5.3.1 Computational expansion of the noscapine pathway

A biosynthetic pathway to a product of interest can potentially be employed to produce numerous derivative compounds by performing chemical transformations on the functional groups of the product and its intermediates. The more complex the molecule (and thus the more functional groups available for derivatization), the more intermediate compounds examined, and the more chemical transformations examined, the larger the number of derivatives that can potentially be made. Furthermore, by iterating this procedure on the products of each chemical transformation, the potential space is further expanded; with each successive generation employed, the number of products increases exponentially. Prior to doing any experiments, the hypothetical space of pathway derivatives can be predicted using reaction prediction tools such as BNICE.ch.

Here, we applied the BNICE.ch network generation on the noscapine pathway, starting from (*S*)-norcoclaurine, using 442 reaction rules. The portion of the noscapine pathway reconstructed with BNICE.ch in this work involves a total of 17 metabolites connected by 17 reactions which are catalyzed by a total of 11 generalized reaction rules. We performed the BNICE.ch network generation starting from all 17 of these metabolites, and initially expanded the pathway for four generations, allowing both known and novel reactions to be generated, as well as allowing all compounds known to any biological, bioactive, or chemical database. This initial expansion produced a network spanning 4,838 compounds and 17,597 reactions (Table 5.6). Of these 17,597 reactions, 244 were known either in the KEGG database or as a part of the noscapine pathway, meaning that each is known to be catalyzed by a well-characterized enzyme that is linked to a genetic sequence from at least one organism. Of the 4,838 compounds, 720 were classified as biological or bioactive, meaning they were found in a biological database or one of the bioactive databases ChEBI or ChEMBL. The remaining compounds are classified as chemical compounds, as information regarding them could only be found in PubChem.

Table 5.6: Overview on BNICE.ch network statistics

		Reactions	Compounds
Raw BNICE.ch network	Total	17597	4838
	Biological / bioactive	244	720
Benzylisoquinoline alkaloid network	Total	7527	1518
	Biological / bioactive	49	99
Potential targets			1501
Potential targets with annotation	(min 1 citation/patent)		545

As our analysis was focused on BIAs, we required at least one compound on each side of the reaction to have a minimal elemental composition of the minimal BIA, 1-benzylisoquinoline, which has 16 carbon atoms, 13 hydrogen atoms, and 1 nitrogen atom. The resultant BIA network spanned 1,518 compounds, of which 99 were known to biological or bioactive databases, and 7,527 reactions, of which 49 were known biological reactions. It was apparent that our network was not uniform about the entire length of the noscapine biosynthetic pathway (Figure 5.9). The upstream portion of the network, nearest to (*S*)-norcoclaurine, is highly connected, whereas the network in the downstream portion closer to noscapine is less populated. This is likely due to the fact that the downstream intermediates in the pathway and their derivatives are increasing in size and complexity, which makes it difficult to detect them experimentally and to characterize their structure. As a consequence, they are less likely to be represented in any biological and chemical database, and therefore not part of the predicted network despite their increased diversity in functional groups.

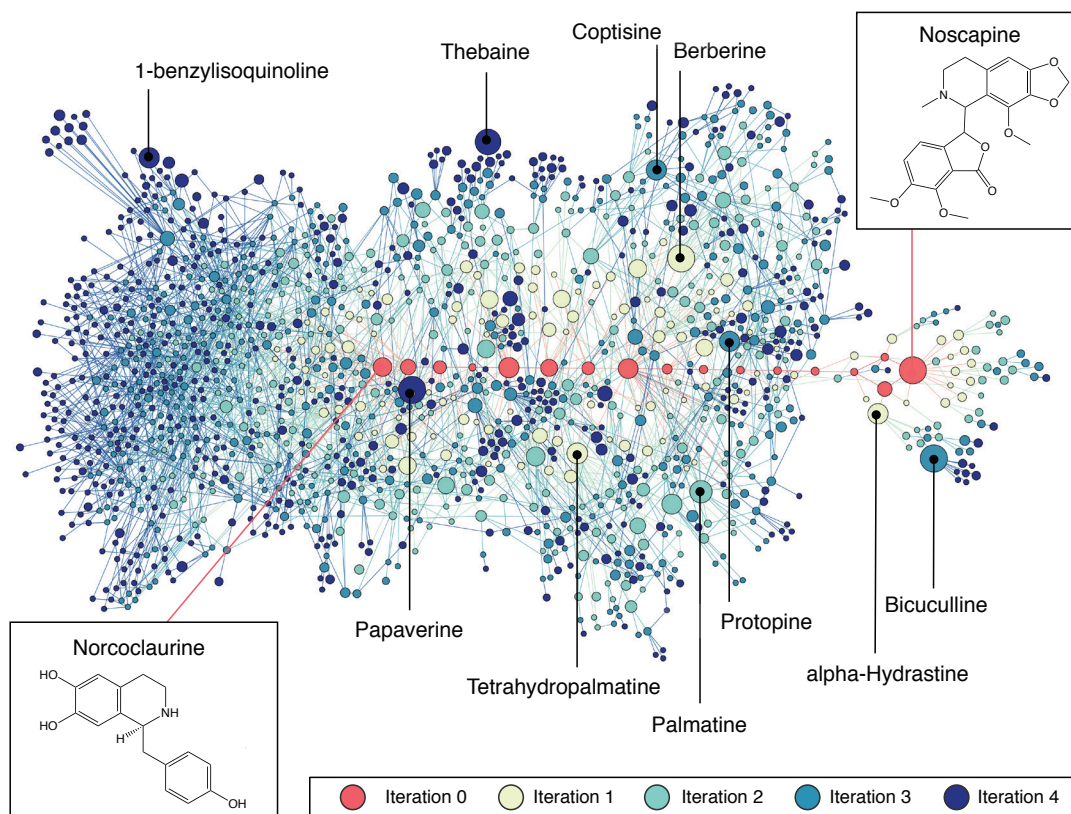


Figure 5.9: Visualization of the expanded biosynthesis network of the noscapine pathway. The nodes and edges drawn in red shows the original noscapine pathway. Around the original pathway, the predicted network of compounds (nodes) and reactions (edges) is visualized. The top 10 compounds in terms of 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 has been 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. The free graph visualization tool Gephi was used for network visualization⁷⁸.

5.3.2 Ranking candidate compounds by popularity

When far more potential compounds are generated by BNICE.ch than can be reasonably experimentally evaluated, as was the case with our exploration of the space surrounding the noscapine pathway, it is necessary to rank the candidate compounds in some manner to guide experimental effort. Numerous ranking criteria could be employed, depending on what properties the desired compound(s) should possess⁷⁹. For example, if searching for new drug candidates, Lipinski's rule of five⁸⁰ could be employed, prioritizing compounds of a given molecular mass, calculated partition coefficient, and/or number of hydrogen bond donors and acceptors. One could also prioritize the chemical novelty of the potential compounds by prioritizing those that have never before been synthesized, in order to most effectively leverage the biosynthesis platform to manufacture molecules that could not be made chemically.

We chose to rank the noscapine-derived compounds by "popularity", a measure of the number of publications and patent annotations, in order to focus on compounds that could be applicable to the work of other researchers; the number of publications was derived

from both 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/>)⁸¹ on the number of associated patents and citations. Complementary, we used the Entrez Programming Utilities (E-utilities) API service to search the PubMed database for citations by compound name⁸². We then screened the 1,501 potential target compounds (1,518 satisfying the BIA requirement described above, minus the 17 compounds that are a part of the noscapine pathway) and found that 204 returned at least one PubMed reference, while 467 had at least one patent associated with them. In total, at least one annotation (citation or patent) was obtained for 545 distinct compounds (Appendix Table A3).

The most popular compound by total rank (the sum of the number of citations and patents) was papaverine, with 22,918 total annotations. The compounds bicuculline and berberine ranked second and third with 16,118 and 12,154 total annotations, respectively. Bicuculline was the most popular compound when ranked specifically by citations, with 13,209 PubMed citations, followed by papaverine with 7,947 and berberine with 5,403. Conversely, papaverine had the highest number of associated patents with 14,971, thus giving it the highest total annotations, followed by berberine with 6,751 patents and thebaine with 4,012 patents. The disparity in ranking between citations and patents observed with bicuculline potentially reflects its research history. While bicuculline is widely employed in medical research to mimic epilepsy in mice, and as such ranks first in number of citations, it is ranked fourth in patent count, possibly reflecting a relative lack of clinical applications. Instead of focusing on total annotations as we have done, the ranking could potentially be restricted to citations, to emphasize research interest, or patents, to emphasize commercial interest. All of the compounds with at least one annotation were considered as potential bioengineering targets.

5.3.3 Construction of biosynthetic pathways to target compounds

While the application of a ranking algorithm to the potential compounds generated by BNICE.ch, whether it prioritizes popularity, novelty, drug-likeness, or some other property, will help to identify top candidate compounds, it will not necessarily prioritize those which can be feasibly produced experimentally. In order to avoid the laborious construction and evaluation of pathways that are unlikely to produce the compound of interest, it is helpful to apply additional computational filters to determine the best candidates for bioproduction. We chose to apply the three following criteria in order to determine a single compound to focus on: (i) the production pathways toward the target is feasible in terms of thermodynamics and enzyme availability; (ii) the target compound is as close as possible to the original pathway in order to minimize the number of enzymes that must be integrated for experimental validation; and (iii) the target molecule is a potential or confirmed pharmaceutical.

We first wanted to ensure the biological feasibility of the potential pathways to our target compounds. For the top 50 ranked potential target compounds, we enumerated all the

possible pathways connecting each target to intermediates in the noscapine pathway within a maximum of four reaction steps using NICEpath. Reactions known to have a high standard Gibbs free energy of reaction, *i.e.* reactions producing molecular oxygen or binding carbon dioxide to the substrate, as well as reactions that demethylate the substrate via *S*-adenosylmethionine were not considered to avoid thermodynamic and catalytic bottlenecks. We found feasible pathways for 42 out of 50 targets, furnishing a total of 1,338 pathways (Table 5.7). To assess the catalytic feasibility of the pathways, we predicted enzymes for each novel reaction step involved using BridgIT¹⁶. BridgIT calculates a reactive-site centric similarity score (BridgIT score) between the novel reaction and a reference database of known, well-characterized reactions by comparing the molecular fingerprints on and around the reactive sites of the reactants. The output is a ranked list of candidate enzymes and associated similarity scores that indicate the probability of the candidate enzyme to catalyze the novel reaction. We collected the best BridgIT hits of each reaction in the pathway and calculated the mean, thus providing an overall metric to describe the catalytic feasibility of the pathways.

Interestingly, the known biosynthetic pathway for protopine and the proposed pathways for papaverine were both part of our solution space, which lends credence to our approach. Furthermore, our solution space suggests additional alternative pathways for the biosynthesis of these compounds. All of our proposed pathways are available in a visualization online at <https://lcsb-databases.epfl.ch/pathways/GraphList>.

We next wanted to ensure the target compound is as close to the original pathway as possible. Given the richness of our reaction network, we were able to restrict our search to compounds only one step away from a noscapine pathway intermediate and still have 15 potential targets, each of them produced by a feasible reaction and associated with a ranked list of predicted, putative enzymes (

Table 5.8). Among these compounds, (*S*)-tetrahydropalmatine was chosen for its high popularity score, its feasible biosynthetic pathway, and its reported medicinal interest. (*S*)-Tetrahydropalmatine is naturally found in plants from the genus *Corydalis* in the Papaveraceae family, and it is also produced by additional plants, *e.g.* *Stephania rotunda*, that are traditionally used in Chinese herbal medicine. (*S*)-Tetrahydropalmatine (synonymous with levotetrahydropalmatine) has been widely used for its analgesic, anxiolytic, and sedative effects as an alternative to opiates and benzodiazepines, and furthermore has shown promise in treating opiate, cocaine, and methamphetamine addiction⁷⁷.

Table 5.7: List of 50 most popular compounds in the generated network, based on number of associated patents and PubMed citations in the PubChem database. The compounds are ranked by the sum of the number of citations and patents. For each potential target, the number of feasible pathways as well as the noscapine pathway precursor are indicated.

Rank	Name	# Citations	# Patents	Total (Citations + Patents)	# Pathways	Precursor
1	Papaverine	7947	14971	22918	204	(S)-Reticuline
2	Bicuculline	13203	2915	16118	0	
3	Berberine	5403	6751	12154	2	(S)-Canadine
4	Thebaine	532	4012	4544	12	(S)-Reticuline
5	Palmitine	892	548	1440	4	Tetrahydrocolumbamine
6	Tetrahydropalmatine	530	355	885	2	Tetrahydrocolumbamine
7	1-benzylisoquinoline	198	602	800	0	
8	Coptisine	451	327	778	85	(S)-Scoulerine
9	Protopine	357	419	776	30	(S)-cis-N-Methylcanadine
10	alpha-Hydrastine	131	623	754	0	
11	Jatrorrhizine	477	233	710	0	
12	Laudanosine	228	283	511	10	(S)-Reticuline
13	Columbamine	131	235	366	1	Tetrahydrocolumbamine
14	Magnoflorine	204	158	362	6	(S)-Reticuline
15	Salutaridine	85	264	349	1	(S)-Reticuline
16	Norlaudanosoline	144	177	321	3	(S)-Norcoclaurine
17	Stepholidine	157	140	297	1	Tetrahydrocolumbamine
18	Allocryptopine	111	159	270	2	(S)-cis-N-Methylcanadine
19	Spinosine	3	262	262	2	3'-Hydroxy-N-methyl-(S)-coclaurine
20	Corydaline	67	117	184	4	Tetrahydrocolumbamine
21	Stylopine	56	116	172	12	(S)-Scoulerine
22	Chileninone	85	70	155	5	(S)-Scoulerine
23	Phellodendrine	42	111	153	3	(S)-Reticuline
24	Laudanidine	23	112	135	3	(S)-Reticuline
25	Cryptopine	28	103	131	420	(S)-Scoulerine
26	Epiberberine	128	0	128	64	(S)-Scoulerine
27	Isocorydine	74	53	127	24	(S)-Reticuline
28	Salutaridinol	30	86	116	2	(S)-Reticuline
29	Tetrahydrobenzylisoquinoline	36	78	114	0	
30	Tetrahydropapaverine	21	89	110	129	(S)-Reticuline
31	1-[(4-methoxyphenyl)methyl]isoquinoline	44	30	74	0	
32	Codamine	13	61	74	2	(S)-Reticuline
33	Norreticuline	33	40	73	3	(S)-Reticuline
34	Dehydrocorydaline	66	6	72	12	Tetrahydrocolumbamine
35	Berlambine	17	53	70	8	(S)-Canadine
36	Corydine	24	34	58	24	(S)-Reticuline
37	Demethyleneberberine	23	34	57	2	Tetrahydrocolumbamine
38	Corytuberine	18	39	57	1	(S)-Reticuline
39	Lambertine	30	23	53	6	(S)-Canadine
40	Xylopinine	16	37	53	43	(S)-Reticuline
41	Corypalmine	11	42	53	0	
42	(1R)-1-(3-Hydroxy-4-methoxybenzyl)-6-methoxy-1,2,3,4-tetrahydroisoquinoline	0	45	45	0	
43	Armepavine	28	15	43	2	(S)-N-Methylcoclaurine
44	1,2-Dehydroreticuline	3	40	43	1	(S)-Reticuline
45	1-benzyl-6,7-dimethoxy-1,2,3,4-tetrahydroisoquinoline	0	41	41	0	
46	Nandinine	1	39	40	1	(S)-Scoulerine
47	N-(5-oxo-4,4-dipropyl-3-oxolanyl)-4-(trifluoromethyl)benzenesulfonamide	3	35	38	0	
48	6,7-dimethoxy-1-[(4-methoxyphenyl)methyl]-1,2,3,4-tetrahydroisoquinoline	0	38	38	18	(S)-Coclaurine
49	4-(1-isoquinolylmethyl)phenol	6	29	35	0	
50	Capaurine	6	29	35	4	Tetrahydrocolumbamine

Table 5.8: List of compounds ordered by descending popularity that are one reaction step away from the noscapine pathway.

Popularity rank	Name	Best BridgIT score	Predicted EC
3	Berberine	1.00	1.3.3.8
6	Tetrahydropalmatine	1.00	2.1.1.89
13	Columbamine	0.99	1.3.3.8
15	Salutaridine	1.00	1.14.19.67
16	Norlaudanoline	0.99	1.14.14.102
17	Stepholidine	0.78	1.14.13.31
18	Allocryptopine	0.32	1.14.13.239
24	Laudanidine	1.00	2.1.1.291
31	Codamine	0.79	2.1.1.121
33	Norreticuline	0.09	1.5.3.10
37	Corytuberine	0.56	1.14.19.67
39	Lambertine	0.45	1.3.1.29
43	Armeopavine	1.00	2.1.1.291
43	1,2-Dehydroreticuline	1.00	1.5.1.27
46	Nandinine	1.00	1.14.19.73

5.3.4 Selection of candidate enzymes for tetrahydropalmatine bioproduction

Once a compound of interest is chosen, there still remains the task of selecting the enzyme or enzymes that are most likely to carry out the desired transformation. While at this point only a single strain is required to evaluate each enzyme candidate, each of which can be synthesized, cloned, and expressed on a plasmid relatively quickly and inexpensively, there are initially an unwieldy number of potential enzyme candidates to evaluate. Since at this step an enzyme candidate is simply any enzyme that is known to carry out the same transformation – to use our example of the desired conversion of (*S*)-tetrahydrocolumbamine to (*S*)-tetrahydropalmatine, any enzyme which performs an O-methylation, regardless of the substrate on which it performs it, is a candidate – the list of candidates can be hundreds or thousands of enzymes long. It is thus useful to perform one last computational ranking before experimental evaluation, this time to determine which of the candidate enzymes are most likely to perform the desired transformation.

We sought to do this by determining which enzymes had native substrates that most closely resembled our desired substrate, (*S*)-tetrahydrocolumbamine, by closely analyzing the BridgIT results obtained in the previous step. (*S*)-Tetrahydropalmatine can be produced in one step via the methylation of the 2-hydroxyl of the noscapine intermediate (*S*)-tetrahydrocolumbamine with the concomitant conversion of S-adenosylmethionine to S-adenosylhomocysteine (Figure 5.10). While this reaction has been reported as a side reaction of some enzymes,⁸³ it is not the native function of those enzymes. As such the KEGG database contained the reaction but without any associated gene sequence, as it mainly focuses on native enzyme functions, and our BNICE.ch search compared its predictions only to the native activity of enzymes reported in KEGG. Hence, BridgIT analysis suggested a ranked list of

enzymes to catalyze this reaction, based on the structural similarity of the (*S*)-tetrahydrocolumbamine methylation to the native reactions of those enzymes (Table 5.9).

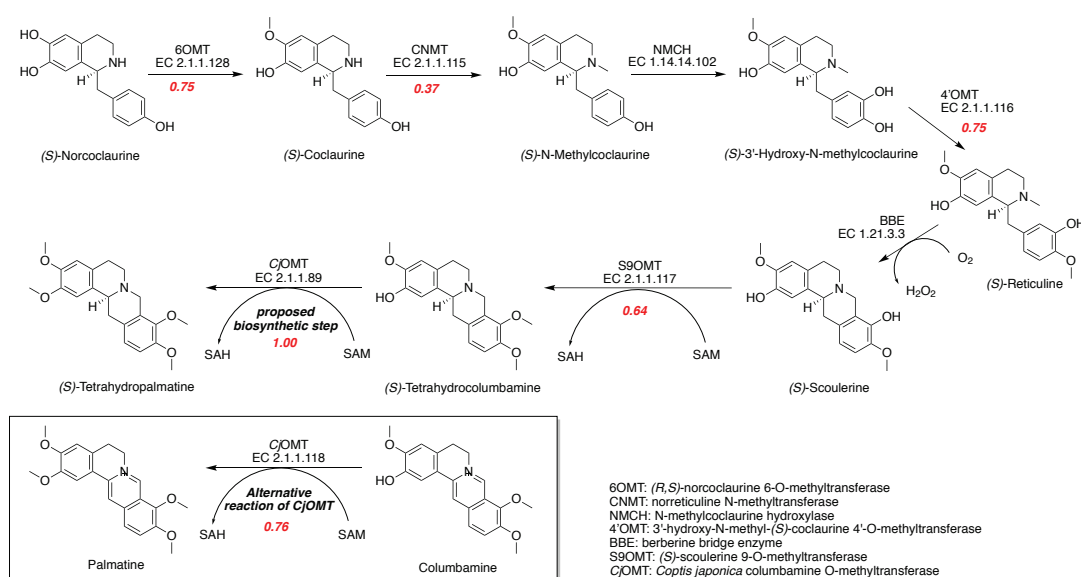


Figure 5.10: Metabolic pathway to (*S*)-tetrahydropalmatine from (*S*)- norcoclaurine in yeast. For each reaction, the enzyme identifier and the EC number are indicated. In addition, the similarity of each reaction with respect to the proposed biosynthetic step for tetrahydropalmatine is indicated with the BridgIT score (bold red) obtained by BridgIT analysis.

Perhaps unsurprisingly, the methyltransferases in the noscapine pathway were found to be high ranking candidates for catalyzing the predicted reaction, as their native substrates are necessarily quite similar in structure. Ranked fourth and fifth, both with BridgIT scores of 0.75, were the enzymes 6OMT (2.1.1.128) and 4'OMT (2.1.1.116), which *O*-methylate the noscapine pathway intermediates (*S*)-norcoclaurine and (*S*)-3'-Hydroxy-N-methylcoclaurine, respectively. The enzyme ranked 15th with a BridgIT score of 0.64 was found to be S9OMT (2.1.1.117), one example of which is *yPsS9OMT*, which was present in our pathway to catalyze the step immediately upstream of our desired transformation. These high BridgIT scores indicated a potential promiscuous activity of these OMTs on (*S*)-tetrahydrocolumbamine. On the other hand, the first native enzyme from *S. cerevisiae*, the poly-prenyldihydroxybenzoate methyltransferase (2.1.1.114), was ranked 24th with a score of 0.59, indicating that there would likely be no interference from native yeast enzymes with the pathway.

The candidate enzyme with the highest BridgIT score was found to be the tetrahydrocolumbamine 2-O-methyltransferase (2.1.1.89) with the maximum possible score of 1.00, indicating identity of query and proposed reaction. However, it turned out to have no sequence assigned in our reference database. One of the highest ranked enzymes was the columbamine O-methyltransferase from *Coptis japonica* (2.1.1.118; referred to here as *CjCo*-OMT), which converts the compound (*S*)-columbamine to (*S*)-palmatine. (*S*)-Columbamine is very similar in structure to (*S*)-tetrahydrocolumbamine, differing only in that the tetrahydroisoquinoline moiety of (*S*)-tetrahydrocolumbamine has effectively undergone a four-electron oxidation to yield the quinoline moiety of (*S*)-columbamine, and as such this was a

very promising candidate for our desired transformation, with a BridgIT score of 0.76. This enzyme had previously been tested *in vitro* and found to possess promiscuous activity on (*S*)-tetrahydrocolumbamine as well. This activity, however, was not linked to any known enzyme in KEGG.

5.3.5 *In vitro* and *in vivo* bioproduction of tetrahydropalmatine

The workflow described in the previous sections served to create a ranked list of enzymes likely to produce the desired product of interest. Many researchers utilizing this workflow might only screen a single-digit number of enzymes to find out that suits their needs; we thus decided to focus our efforts on a similarly small selection in order to mirror future users' experiences. While the primary concern is to validate that the candidate enzyme successfully produces the product of interest *in vivo* when integrated into the parent biosynthetic pathway, further characterization of the candidate enzyme is often beneficial. This is particularly true when high titers of the desired product are required, as enzyme engineering may need to be performed on the candidate enzyme to increase its activity, in particular in case the substrate is non-native.

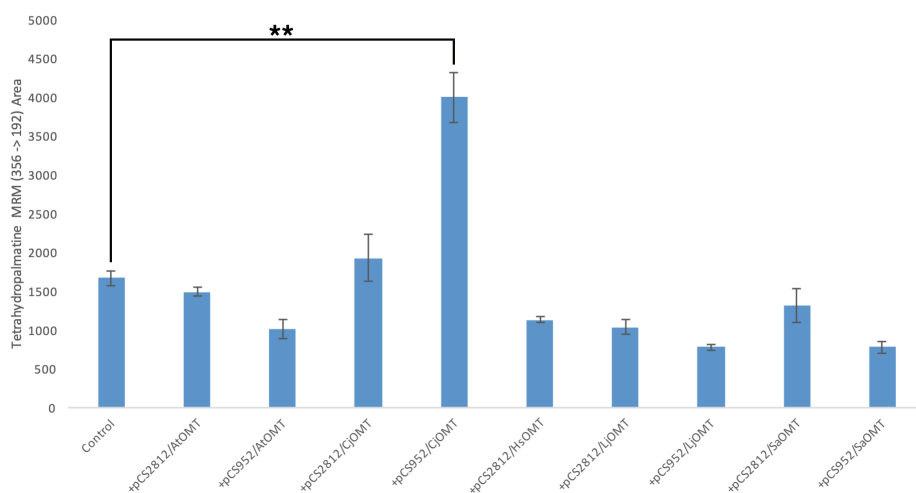
In order to survey an experimentally manageable but still diverse set of enzymes, six of the top 25 hits from BridgIT were chosen for experimental validation. These enzymes were selected due to the broad range of native substrates they possess as well as their diverse parent organisms, with hits screened that are of plant, bacterial, and animal origin (Table 5.9). These six candidates – *CjColOMT*, *ObEugOMT*, *AtCafOMT*, *HsSerOMT*, *SaPurOMT*, and *LjFlaOMT* – were expressed off of plasmids in the yeast strain YCS1171, a *de novo* (*S*)-reticuline biosynthetic strain which synthesizes (*S*)-reticuline from fed sugars⁷⁴. The predicted OMTs were each cloned into two *S. cerevisiae* expression vectors, one plasmid containing a high-copy number origin of replication, the other containing a low-copy number. One of the OMTs, *ObEugOMT*, failed to clone into either vector, while one other, *HsSerOMT*, cloned into only the low-copy construct. The other nine constructs were all cloned successfully, verified by sequencing, and used to transform the *de novo* THCB producing strain. In order to detect the (*S*)-tetrahydropalmatine potentially produced by the *in silico* predicted enzymes, a commercial standard of (*S*)-tetrahydropalmatine was purchased and first used to develop an optimized LC-MS/MS method. After three days' growth in defined media, the strains were analyzed for (*S*)-tetrahydropalmatine production.

The highest ranked candidate, *CjColOMT*, was found to produce a statistically significant increase in the amount of (*S*)-tetrahydropalmatine produced (Figure 5.11). Remarkably, we also observed the production of a lower amount of (*S*)-tetrahydropalmatine in every strain tested. Our first suspicion was that one or more of the other methyltransferases present in the (*S*)-tetrahydrocolumbamine pathway was responsible for the production of the background level of (*S*)-tetrahydropalmatine; the native substrates of these enzymes are precursors of (*S*)-tetrahydrocolumbamine that are very similar in structure to (*S*)-tetrahydrocolumbamine, and thus it is possible that they may be able to accommodate the chemically related (*S*)-tetrahydrocolumbamine as well. In fact, three of the methyltransferases in the

pathway – S9OMT, which acts natively on (*S*)-scoulerine, 6OMT, which acts on norcoclaurine, and 4'OMT, which acts on 6-methyl-(*S*)-laudanosoline – were assigned high scores by the BridgIT analysis for their potential activity on (*S*)-tetrahydrocolumbamine.

Table 5.9: Reaction similarities between the predicted tetrahydropalmatine-producing reaction and its top 25 most similar reactions from the BridgIT reference database.

Rank	BridgIT score	Predicted EC	Enzyme	Native substrate	Type of substrate	Native organism	Activity on THCB
1	1.00	2.1.1.89					Predicted reaction ¹
2	0.98	2.1.1.291	Ps7OMT	(<i>S</i>)-Reticuline	BIA	<i>P. somniferum</i>	Not tested
3	0.76	2.1.1.118	C/ColOMT ²	Columbamine	BIA	<i>Coptis japonica</i>	Active
4	0.75	2.1.1.128	Ps6OMT	(<i>S</i>)-Norcoclaurine	BIA	<i>P. somniferum</i>	Already in pathway
5	0.75	2.1.1.116	Ps4'OMT	3'-Hydroxy-N-methyl-(<i>S</i>)-coclaurine	BIA	<i>P. somniferum</i>	Already in pathway
6	0.73	2.1.1.121					No enzyme available
7	0.72	2.1.1.146	ObEugOMT	Isoeugenol	Phenylpropanoid	<i>Ocimum basilicum</i> (Basil)	No activity
8	0.72	2.1.1.323					No enzyme available
9	0.69	2.1.1.330					No enzyme available
10	0.69	2.1.1.38	SaPurOMT	O-Demethylpuromycin	Antibiotic	<i>Streptomyces alboniger</i>	No activity
11	0.68	2.1.1.6	CatOMT	Catechol	Phenol	Diverse	Not tested
12	0.66	2.1.1.212	LjFlaOMT	2,4',7-Trihydroxyisoflavanone	Flavanonoid	<i>Lotus japonica</i>	No activity
13	0.65	2.1.1.336					No enzyme available
14	0.64	2.1.1.4	HsSerOMT	N-Acetylserotonin	Neurotransmitter	<i>Homo sapiens</i> (Human)	No activity
15	0.64	2.1.1.117	yPsS9OMT ³	(<i>S</i>)-Scoulerine	BIA	<i>P. somniferum</i>	Already in pathway, active
16	0.63	2.1.1.231	Fla4'OMT	4'-Hydroxyflavone	Flavonoid	<i>Glycine max</i> (Soybean)	Not tested
17	0.63	2.1.1.68	Caf3OMT	(<i>E</i>)-Caffeate	Phenylpropanoid	Diverse	Not tested
18	0.62	2.1.1.104	AtCafOMT	Caffeoyl-CoA	Phenylpropanoid	<i>Arabidopsis thaliana</i>	No activity
19	0.61	2.1.1.150	Caf3OMT	(<i>E</i>)-Caffeate	Phenylpropanoid	<i>Medicago sativa</i> (Alfalfa)	Not tested
20	0.61	2.1.1.222	UbOMT	3-Demethylubiquinol	Quinone	Diverse bacteria	Not tested
21	0.61	2.1.1.108					No enzyme available
22	0.60	2.1.1.279	AnOMT	trans-Anol	Phenol	<i>Pimpinella anisum</i> (Anise)	Not tested
23	0.60	2.1.1.94	Tab16OMT	16-Hydroxytabersonine	Terpene indole alkaloid	<i>Catharanthus roseus</i>	Not tested
24	0.59	2.1.1.114		3,4-Dihydroxy-5- <i>all-trans</i> -polyprenylbenzoate		Diverse, incl. <i>S. cerevisiae</i>	Natively present yeast
25	0.59	2.1.1.25					No enzyme available



** = $p < 0.01$

Figure 5.11: Tetrahydropalmatine production in yeast transformed with different OMT-encoding plasmids after 5 days growth. +pCS2812: low-copy number plasmid, +pC2952: high-copy number plasmid. MRM: Multiple Reaction Monitoring counts.

In addition to S9OMT, 6OMT, and 4'OMT, there is one additional methyltransferase present in the (*S*)-tetrahydrocolumbamine pathway: CNMT, which acts natively on 6-methyl-(*S*)-nornocclaurine. In the pathway we originally constructed, the specific isoforms of these four methyltransferases were all derived from *Papaver somniferum*, and thus were specifically named *Ps*6OMT, *Ps*CNMT, *Ps*4'OMT, and *yPs*S9OMT (the *y* in the last enzyme denoting that it had been codon optimized for *S. cerevisiae*). When attempting to purify these enzymes from *E. coli*, conditions were found that furnished *Ps*6OMT, *Ps*CNMT, and *yPs*S9OMT, as well as *Cj*ColOMT, but no conditions could be found that afforded soluble *Ps*4'OMT. Accordingly, we then looked to alternative isoforms of 4'OMT from other species, and chose three – *Cj*4'OMT, *Ec*4'OMT, and *Tf*4'OMT – of which we integrated five variants (both *E. coli* codon optimized and *S. cerevisiae* codon optimized versions of *Cj*4'OMT and *Ec*4'OMT were used) into the *de novo* (*S*)-tetrahydrocolumbamine strain to replace the copies of *Ps*4'OMT present. All five of these strains still produced high titers of (*S*)-tetrahydrocolumbamine, and all five still displayed increased production of (*S*)-tetrahydropalmatine when *Cj*ColOMT was expressed on a plasmid.

All five of the alternative 4'OMTs were found to express well in *E. coli* and were purified for *in vitro* analysis. The substrate for the reaction, (*S*)-tetrahydrocolumbamine, was generated from a large-scale *in vitro* bioconversion of (*S*)-scoulerine, which is commercially available, with a variant of *Tf*S9OMT. *In vitro* bioconversions of (*S*)-tetrahydrocolumbamine were then performed with purified stocks of either *Ps*6OMT, *Ps*CNMT, one of the five 4'OMTs, *yPs*S9OMT, and *Cj*ColOMT. These reactions showed no *in vitro* production of (*S*)-tetrahydropalmatine by *Ps*6OMT, *Ps*CNMT, or any of the 4'OMTs, and reactions on their native substrates confirmed that the purified enzymes obtained are all active, but simply possess no activity on the non-native substrate (*S*)-tetrahydrocolumbamine (Figure 5.12A). It is worth noting that *Ps*CNMT did accept (*S*)-tetrahydrocolumbamine as a substrate, but the resultant product was presumably the *N*-methylated derivative, as no (*S*)-tetrahydropalmatine production was observed with *Ps*CNMT (Figure 5.12B). In contrast, a small amount of (*S*)-tetrahydropalmatine was observed to be produced by *yPs*S9OMT, while a significantly larger amount was produced by *Cj*ColOMT. The fact that these two enzymes were highly ranked candidates produced by our computational workflow, with *Cj*ColOMT having the second highest BridgIT score of all enzymes in the reference reaction database, validates our workflow for predicting new products accessible from a biosynthetic pathway and providing actionable, effective enzyme suggestions. From one perspective, the success of these two enzymes may seem unsurprising; both perform similar reactions on chemically very similar substrates. In fact, both of these enzymes have been reported to have promiscuous activity toward (*S*)-tetrahydrocolumbamine *in vitro*⁸⁴ but again, non-native enzyme activities were not initially considered by ourselves, BNICE.ch, or BridgIT when making predictions. In cases where such non-native activity data are available, they may be buried in the literature and not a part of an easily searchable database, and thus might be overlooked by or unavailable to bioengineers; in these cases, our workflow can rapidly provide predictions which recapitulate these data. And in cases where such data are not known, our

workflow has demonstrated that it is capable of inferring likely off-target activity from only native enzyme data.

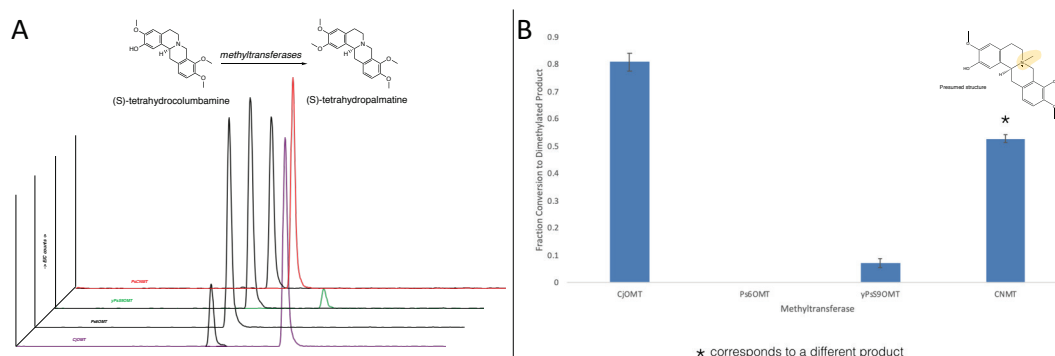


Figure 5.12: (A) Screening methyltransferases in tetrahydropalmatine pathway for enzymatic activity on (S)-tetrahydrocolumbamine. (B) Quantification of enzyme activity, comparing CjOMT, Ps6OMT, yPsS9OMT, CNMT. The product of CNMT is not tetrahydropalmatine, but presumably the N-methylated product of tetrahydrocolumbamine (structure shown). EIC: Extracted Ion Counts.

5.3.6 Conclusion

We developed a new pipeline to specifically explore the biochemical surroundings of biosynthetic pathways for PNPs and we successfully predicted and implemented the production of a pharmaceutical molecule, (S)-tetrahydropalmatine, in a host organism suited for industrial production. We further confirmed our approach by recovering known biosynthesis pathways for papaverine and protopine, and we also suggest new biosynthetic routes for these and other compounds of potential pharmaceutical value. Our workflow was able to predict promiscuous enzymatic activities of two of the top candidates, CjColOMT and yPsS9OMT, and we confirmed the activity of these enzymes on (S)-tetrahydrocolumbamine *in vitro*. We also discovered the N-methylating activity of PsCNMT on tetrahydrocolumbamine, leading to the production of N-methylated THCB. Our workflow can be applied to any other class of natural products to (i) expand the bioproduction scope of existing producer strains (for example for opioids⁷³, flavonoids^{85,86}, cannabinoids⁸⁷, carotenoids⁸⁸) or (ii) to generate hypotheses about the presence of potential side-products of biosynthetic pathways that have previously gone undetected. The biochemical network presented in this work is a collection of bioengineering opportunities for scientists aiming to produce one or several of the detected metabolites in the vicinity of noscapine. We believe that our approach has the power to direct research towards new discoveries and to drive engineering efforts towards the bioproduction of valuable chemicals and pharmaceuticals.

5.4 Predicting potential biodegradation of xenobiotics

The work presented in this Subchapter is the result of a master project accomplished by Basile Laurent, supervised by the author of the thesis. The purpose of this Subchapter is illustrative since the main work has been presented as a master thesis.

Xenobiotics are man-made chemicals that do not naturally occur in organisms, and for many of them nature does not have the enzymatic mechanisms in place to break them down into the basic building blocks of life. Anthropogenic chemicals can be a danger to humans and to the ecosystem in general if they are toxic to living organisms, or if they are released in large amounts into nature, as it is the case for PET. A broader definition of xenobiotics also includes naturally occurring compounds (*e.g.*, antibiotics) that are released into the environment in big quantities where they threaten the natural ecosystem.

Microorganisms can be harnessed to degrade pollutants in wastewater treatment, or in the bioremediation of chemically polluted sites. However, for many xenobiotics there is currently no biological strategy to remove them from the environment. To amend this, computational tools have been developed in the past to systematically study biodegradation pathways, and to predict potential biodegradation routes. One important tool is *enviPath*⁵⁰, an online resource that collects biodegradation pathways and provides predictions based on generalized reaction rules specific to biodegradation reaction mechanisms. BNICE.ch has also been applied in the past to predict biodegradation pathways for 4-chlorobiphenyl, phenanthrene, γ -hexachlorocyclohexane, and 1,2,4-trichlorobenzene⁸⁹. More general, the applications of systems biology to the biodegradation and bioremediation problem have been comprehensively reviewed by de Lorenzo in 2008, and Dvořák et al. in 2017^{90,91}. One of the discussed ideas is the application of genetic engineering to create strains capable of degrading problematic pollutants⁹².

However, there are several problematic aspects in the application of genetically modified organisms in bioremediation and wastewater treatment⁹³. First of all, the release of genetically modified organisms into the environment obstacle can be problematic because of safety concerns and regulations. Second, engineered organisms are generally more vulnerable than their wild-type counterparts, which can be due to an increased plasmid burden, or to a suboptimal redistribution of metabolic resources that decreases the overall robustness of the host. To bypass active genetic engineering of microbes, we hypothesized that there might exist organisms in nature that already have a native capacity to degrade a given pollutant. These organisms would use a given compound as a carbon or energy source, or they would have the capability to reroute their metabolism to degrade it. Finally, some organisms might have the right composition of enzymes that would evolve to degrade the pollutant under the pressure of natural selection in the presence of the compound. Unfortunately, many microorganisms are not culturable and therefore difficult to study experimentally. Advances in genome sequencing, however, made available the genomes of many microorganisms found in the environment. This wealth of genetic information can be mined for organisms that have the metabolic capabilities to degrade xenobiotics⁹⁴.

Here, we present a workflow to computationally predict biosynthesis pathways for a given xenobiotic and to find the enzymatically best-suited organisms from available genomics databases. To do this, we find putative enzymes for each reaction step in the predicted biodegradation pathway, and we mine the database UniProtKB⁹⁵ for organisms that have all the necessary enzymes to perform these functions. UniProtKB is a community-driven database of functionally annotated protein sequences collected from whole-genome sequencing data, as well as from specific experimental evidence. It is divided into the high-quality database Swiss-Prot containing manually annotated protein sequences, and TrEMBL, containing automatically annotated proteins based on sequence homology. By taking advantage of the vast collection of data in UniProtKB, we tried to find organisms with the necessary genes to perform the desired reactions. Since many of the predicted reaction steps predicted by BNICE.ch are unknown, the enzymes assigned might not be known to perform the desired functions. In this case, we rely on the metabolic plasticity of prokaryotes to adapt to feed on the new, xenobiotic carbon sources⁹⁶.

5.4.1 Predictive biodegradation workflow

In this study, we wanted to assess if we can adapt the previously described retrobiosynthesis tools to determine the biodegradability of xenobiotics. The proposed workflow consists of four steps (Figure 5.13): First, a hypothetical reaction network is generated around the xenobiotic compound of interest using BNICE.ch, only allowing known chemicals in the network. Second, we search for potential biodegradation routes using NICEpath. To ensure that the end points of the pathways (*i.e.*, the biodegradation products) are known metabolites that can be readily converted by the majority of organisms, we search for pathways that connect to metabolites present in *E. coli*. The reason for this choice is the completeness of the *E. coli* genome-scale model, and the absence of specific biodegradation capabilities and extended secondary metabolism that would bias our results. Also, since we knew in advance that most of the organisms identified from UniProtKB would not have a genome-scale model available for stoichiometric and thermodynamic analysis, we decided to implement thermodynamic constraints (*i.e.*, removing O₂-producing/CO₂-fixing reactions) directly at the pathway search level. In a third step, we assigned putative enzymes and corresponding similarity scores to each reaction step in the pathway using BridgIT. For each identified enzyme with a BridgIT score higher than 0.3 (recommended threshold), we retrieved all the prokaryotes from UniProtKB that had a gene with a corresponding EC annotation. The UniProtKB entries were classified into “reviewed” if they were part of the Swiss-Prot database, and “unreviewed” if they were part of TrEMBL. Finally, we analyzed each of the predicted pathways to see if we could find organisms that have potentially catalyzing enzyme for each reaction step in the pathway. If such an organism could be found, we assigned a pathway score to the combination of pathway and organism that reflects the BridgIT scores, as well as the conserved atom ratios (CARs) of each reaction step involved. More precisely, each reaction was assigned a reaction score which consist of one third of the CAR, plus two thirds of the BridgIT score. The overall pathway score was calculated as the geometric mean of reaction scores, which is the product of reaction scores in the pathway to the power of one

over the length of the pathway. The pathway score allowed us to obtain a ranked list of organisms that might be able to degrade the compound under study.

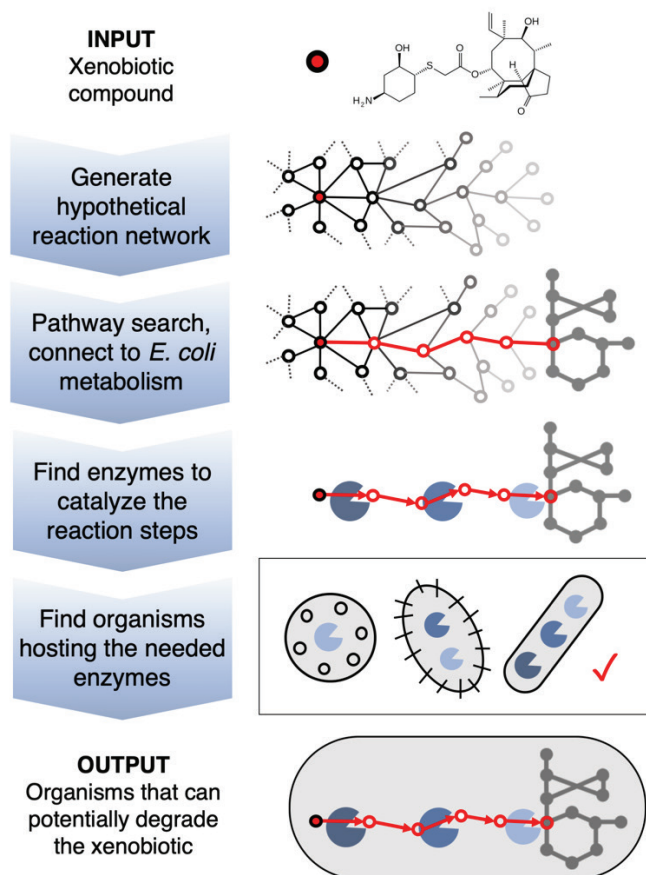


Figure 5.13: Workflow to assess the biodegradability of xenobiotics and to identify organisms with the potential capacity to degrade the xenobiotic of interest.

5.4.2 Evaluation of biodegradation for six xenobiotics

To evaluate the utility of our approach, we applied the workflow on six xenobiotics of different sources and belonging to different chemical classes (Figure 5.14A). The first compound, toluene, has been chosen because it is known to be degraded by several bacteria and fungi, and could therefore be used to validate our workflow. We also analyzed a polyethylene terephthalate (PET) dimer, which is a major threat to marine life. Recently, scientists discovered that the bacterium *Ideonella sakaiensis* could degrade PET thanks to PET-hydrolyzing enzyme, named PETase⁹⁷. The two antibiotics lefamulin and erythromycin were chosen for their chemical complexity, especially the presence of big carbon cycles. The degradation of antibiotics is of major concern in wastewater treatment, since their release into the environment encourages the emergence and spread of resistance mechanisms against these compounds. The list of compounds is completed by the two fungicides isopyrazam and prothioconazole, both environmental pollutants that accumulate in soil and water. The workflow was applied to each of these compounds, and the outcome is discussed in the following.

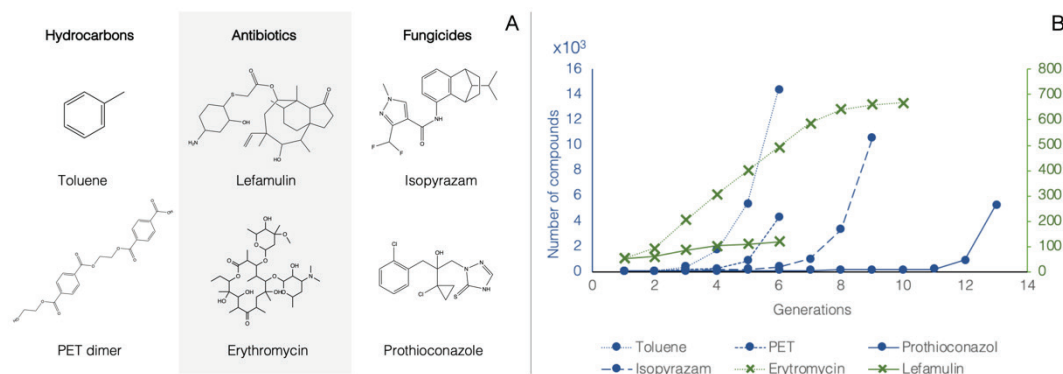


Figure 5.14: (A) Compounds chosen as input for the proposed biodegradation workflow. (B) Network generation around six xenobiotics using BNICE.ch. The number of compounds is indicated after each iteration in the network generation process. Compounds colored in blue map to the left scale, and compounds colored in green map to the right scale.

We were able to generate reaction networks around all of the six compounds within the known compound space (Figure 5.14B), and we observed two types of networks: (i) networks facing a combinatorial explosion of compounds, and (ii) networks reaching a plateau and converge after some generations. The two antibiotics belonged to the second category, while the other four compounds belonged to the first category. For all of the input compounds, we retrieved pathways connecting to *E. coli* metabolisms, annotated the predicted reactions with enzymes using BridgIT, and finally collected the organisms known to express these enzymes. Preliminary results suggest that we could identify organisms for toluene, PET and isopyrazam. For the two antibiotics lefamulin and erythromycin and for prothioconazole however, no organisms could be found. We will first discuss the preliminary results obtained for toluene, PET and isopyrazam, and then discuss the bottlenecks we identified in degradation of the three remaining compounds.

Among the top ten organisms for toluene, we could find *Pseudomonas fluorescens* and *Klebsiella pneumoniae*, which had pathway scores of 0.56 and were associated with toluene degradation in literature^{98,99}. Another organism found in the top ten, *Mycobacterium chlorophenolicum* (also known as *Mycobacterium chlorophenolicum*) with a pathway score of 0.57, is known to degrade the structurally similar compound chlorophenol¹⁰⁰. For two other identified organism, *Cedecea lapagei* and *Microbacterium trichothecenolyticum* with pathway scores of 0.58 and 0.57, respectively, no evidence for toluene degradation could be found in the literature. For the degradation of PET, the preliminary results suggest *Pseudomonas aeruginosa*, *Pseudomonas fluorescens*, *Pseudomonas putida*, *Achromobacter piechaudii* ATCC 43553, and *Cupriavidus oxalaticus*. Interestingly, the three *Pseudomonas* species have been reported to degrade the synthetic polymers nylon, polyester and polyurethane^{101,102}. The known PET-degrader *Ideonella sakaiensis*, however, has not been identified so far in our results, even though its genome is available and annotated in UniProtKB. For the degradation of the fungicide isopyrazam, preliminary results suggest *Cupriavidus taiwanensis* with a pathway score of 0.59. This organism has been reported to degrade the

pesticide chlorpyrifos¹⁰³. The presented examples taken from our preliminary results show that our workflow has indeed the capacity to identify organisms capable of biodegradation for a chemical compound of interest, only given its molecular structure. Based on these first insights, we will be able to fine-tune the parameters used in the workflow, and to further investigate on the obtained results.

For the two antibiotics and for prothioconazole, no organisms could be identified. The network expansion of both lefamulin and erythromycin converged after six and ten generations, respectively. Closer investigations revealed that both compounds could be broken down into smaller structures, but that the main carbon rings (*i.e.*, cyclooctane ring for lefamulin, and erythronolide B for erythromycin) would not undergo further reactions in BNICE.ch. Effectively, none of their potential degradation products was part of any chemical database. Hence, we will need to consider novel compounds in the BNICE.ch network generation to further study the degradation of these substructures. For prothioconazole, all of the predicted pathways included a biotransformation step for which no catalyzing enzyme could be found with a BridgIT score above the set threshold of 0.3. Missing an enzyme for this crucial step, no organisms could be identified. Further investigation into this compound will include allowing novel compounds in the network generation, extended pathway search to find alternative pathways, and the reconsideration of our biochemical reaction rules for biodegradation purpose.

5.4.3 Conclusions

We demonstrated that our workflow has the potential to identify organisms with xenobiotic-degrading capacities, given only the molecular structure of the pollutant. Even though we could not find organisms for all of the compounds studied, we were able to identify potential biochemical bottlenecks in the degradation of these compounds. The presented work marks a first step towards the integration of sequence data into a BNICE.ch-based workflow. However, it still has potential for optimization, in particular regarding the scoring of pathways and organisms. Furthermore, finding branched pathways instead of linear pathways would help us to trace all of the potential degradation products, especially in the case of big, complex compounds such as antibiotics. Finally, our approach can be extended to eukaryotes to identify pollutant degrading fungi and plants, the latter especially for soil bioremediation. We believe that that mining the wealth of available sequencing data is a promising alternative to identify organisms with desired metabolic functions, and to avoid the use of genetic engineering for applications where it is not easily applicable.

References

1. Szymkuć, S. *et al.* Computer-Assisted Synthetic Planning: The End of the Beginning. *Angew. Chemie Int. Ed.* **55**, 5904–5937 (2016).
2. Klucznik, T. *et al.* Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory. *Chem* **4**, 522–532 (2018).
3. Bachmann, B. O. Biosynthesis: Is it time to go retro? *Nat. Chem. Biol.* **6**, 390–393 (2010).
4. Hatzimanikatis, V. *et al.* Exploring the diversity of complex metabolic networks. *Bioinformatics* **21**, 1603–1609 (2005).
5. Tokić, M. *et al.* Discovery and Evaluation of Biosynthetic Pathways for the Production of Five Methyl Ethyl Ketone Precursors. *ACS Synth. Biol.* acssynbio.8b00049 (2018). doi:10.1021/acssynbio.8b00049
6. Delépine, B., Duigou, T., Carbonell, P. & Faulon, J.-L. RetroPath2.0: A retrosynthesis workflow for metabolic engineers. *Metab. Eng.* **45**, 158–170 (2018).
7. Koch, M., Duigou, T. & Faulon, J.-L. Reinforcement Learning for Bio-Retrosynthesis. *bioRxiv* 800474 (2019). doi:10.1101/800474
8. Kumar, A., Wang, L., Ng, C. Y. & Maranas, C. D. Pathway design using de novo steps through uncharted biochemical spaces. *Nat. Commun.* **9**, 184 (2018).
9. Wang, L., Ng, C. Y., Dash, S. & Maranas, C. D. Exploring the combinatorial space of complete pathways to chemicals. *Biochem. Soc. Trans.* **46**, 513–522 (2018).
10. Hadadi, N. & Hatzimanikatis, V. Design of computational retrobiosynthesis tools for the design of de novo synthetic pathways. *Curr. Opin. Chem. Biol.* **28**, 99–104 (2015).
11. Jeffries, J. G., Seaver, S. M. D., Faria, J. P. & Henry, C. S. A pathway for every product? Tools to discover and design plant metabolism. *Plant Sci.* (2018). doi:10.1016/J.PLANTSCI.2018.03.025
12. Lin, G.-M., Warden-Rothman, R. & Voigt, C. A. Retrosynthetic design of metabolic pathways to chemicals not found in nature. *Curr. Opin. Syst. Biol.* (2019). doi:10.1016/J.COISB.2019.04.004
13. Wang, L., Dash, S., Ng, C. Y. & Maranas, C. D. A review of computational tools for design and reconstruction of metabolic pathways. *Synth. Syst. Biotechnol.* **2**, 243–252 (2017).
14. Cravens, A., Payne, J. & Smolke, C. D. Synthetic biology strategies for microbial biosynthesis of plant natural products. *Nat. Commun.* **10**, 2142 (2019).
15. Jankowski, M. D., Henry, C. S., Broadbelt, L. J. & Hatzimanikatis, V. Group Contribution Method for Thermodynamic Analysis of Complex Metabolic Networks. *Biophys. J.* **95**, 1487–1499 (2008).
16. Hadadi, N., MohammadiPeyhani, H., Miskovic, L., Seijo, M. & Hatzimanikatis, V.

- Enzyme annotation for orphan and novel reactions using knowledge of substrate reactive sites. *Proc. Natl. Acad. Sci. U. S. A.* 201818877 (2019). doi:10.1073/pnas.1818877116
17. Hadadi, N., Hafner, J., Shajkofci, A., Zisaki, A. & Hatzimanikatis, V. ATLAS of Biochemistry: A Repository of All Possible Biochemical Reactions for Synthetic Biology and Metabolic Engineering Studies. *ACS Synth. Biol.* (2016). doi:10.1021/acssynbio.6b00054
 18. Hadadi, N. *et al.* A computational framework for integration of lipidomics data into metabolic pathways. *Metab. Eng.* **23**, 1–8 (2014).
 19. Brunk, E., Neri, M., Tavernelli, I., Hatzimanikatis, V. & Rothlisberger, U. Integrating computational methods to retrofit enzymes to synthetic pathways. *Biotechnol. Bioeng.* **109**, 572–582 (2012).
 20. Henry, C. S., Broadbelt, L. J. & Hatzimanikatis, V. Discovery and analysis of novel metabolic pathways for the biosynthesis of industrial chemicals: 3-hydroxypropanoate. *Biotechnol. Bioeng.* n/a-n/a (2010). doi:10.1002/bit.22673
 21. Finley, S. D., Broadbelt, L. J. & Hatzimanikatis, V. Computational framework for predictive biodegradation. *Biotechnol. Bioeng.* **104**, 1086–1097 (2009).
 22. Finley, S. D., Broadbelt, L. J. & Hatzimanikatis, V. Thermodynamic analysis of biodegradation pathways. *Biotechnol. Bioeng.* **103**, 532–541 (2009).
 23. Jeffryes, J. G. *et al.* MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. *J. Cheminform.* **7**, 44 (2015).
 24. Soh, K. C. & Hatzimanikatis, V. DREAMS of metabolism. *Trends Biotechnol.* **28**, 501–508 (2010).
 25. Hatzimanikatis, V., Li, C., Ionita, J. A. & Broadbelt, L. J. Metabolic networks: enzyme function and metabolite structure. *Curr. Opin. Struct. Biol.* **14**, 300–306 (2004).
 26. Yang, X. *et al.* Systematic design and in vitro validation of novel one-carbon assimilation pathways. *Metab. Eng.* **56**, 142–153 (2019).
 27. Mellor, J., Grigoras, I., Carbonell, P. & Faulon, J.-L. Semisupervised Gaussian Process for Automated Enzyme Search. *ACS Synth. Biol.* **5**, 518–528 (2016).
 28. Carbonell, P. *et al.* Selenzyme: enzyme selection tool for pathway design. *Bioinformatics* **34**, 2153–2154 (2018).
 29. Planson, A.-G., Carbonell, P., Paillard, E., Pollet, N. & Faulon, J.-L. Compound toxicity screening and structure-activity relationship modeling in *Escherichia coli*. *Biotechnol. Bioeng.* **109**, 846–850 (2012).
 30. Carbonell, P., Planson, A.-G., Fichera, D. & Faulon, J.-L. A retrosynthetic biology approach to metabolic pathway design for therapeutic production. *BMC Syst. Biol.* **5**, 122 (2011).

31. Carbonell, P., Parutto, P., Herisson, J., Pandit, S. B. & Faulon, J.-L. XTMS: pathway design in an eXTended metabolic space. *Nucleic Acids Res.* **42**, W389–W394 (2014).
32. Carbonell, P., Planson, A.-G. & Faulon, J.-L. Retrosynthetic Design of Heterologous Pathways. in 149–173 (Humana Press, Totowa, NJ, 2013). doi:10.1007/978-1-62703-299-5_9
33. Carbonell, P., Fichera, D., Pandit, S. B. & Faulon, J.-L. Enumerating metabolic pathways for the production of heterologous target chemicals in chassis organisms. *BMC Syst. Biol.* **6**, 10 (2012).
34. Fehér, T. *et al.* Validation of RetroPath, a computer-aided design tool for metabolic pathway engineering. *Biotechnol. J.* **9**, 1446–1457 (2014).
35. Noor, E., Haraldsdóttir, H. S., Milo, R. & Fleming, R. M. T. Consistent Estimation of Gibbs Energy Using Component Contributions. *PLoS Comput. Biol.* **9**, e1003098 (2013).
36. Campodonico, M. A., Andrews, B. A., Asenjo, J. A., Palsson, B. O. & Feist, A. M. Generation of an atlas for commodity chemical production in *Escherichia coli* and a novel pathway prediction algorithm, GEM-Path. *Metab. Eng.* **25**, 140–158 (2014).
37. Mavrovouniotis, M. L. Group contributions for estimating standard gibbs energies of formation of biochemical compounds in aqueous solution. *Biotechnol. Bioeng.* **36**, 1070–1082 (1990).
38. Schilling, C. H., Thakar, R., Travnik, E., Van Dien, S. & Wiback, S. SimPheny™: A Computational Infrastructure for Systems Biology.
39. Yim, H. *et al.* Metabolic engineering of *Escherichia coli* for direct production of 1,4-butanediol. *Nat. Chem. Biol.* **7**, 445–452 (2011).
40. Sivakumar, T. V., Giri, V., Park, J. H., Kim, T. Y. & Bhaduri, A. ReactPRED: A tool to predict and analyze biochemical reactions. *Bioinformatics* btw491 (2016). doi:10.1093/bioinformatics/btw491
41. Tyzack, J. D., Ribeiro, A. J. M., Borkakoti, N. & Thornton, J. M. Exploring Chemical Biosynthetic Design Space with Transform-MinER. *ACS Synth. Biol.* **8**, 2494–2506 (2019).
42. Rodrigo, G., Carrera, J., Prather, K. J. & Jaramillo, A. DESHARKY: automatic design of metabolic pathways for optimal cell growth. *Bioinformatics* **24**, 2554–2556 (2008).
43. Martin, C. H., Nielsen, D. R., Solomon, K. V. & Prather, K. L. J. Synthetic Metabolism: Engineering Biology at the Protein and Pathway Scales. *Chem. Biol.* **16**, 277–286 (2009).
44. Prather, K. L. J. & Martin, C. H. De novo biosynthetic pathways: rational design of microbial chemical factories. *Curr. Opin. Biotechnol.* **19**, 468–474 (2008).
45. Cho, A., Yun, H., Park, J., Lee, S. & Park, S. Prediction of novel synthetic pathways for the production of desired chemicals. *BMC Syst. Biol.* **4**, 35 (2010).

46. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
47. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.* **42**, D459–D471 (2014).
48. Schomburg, I. *et al.* BRENDA: a resource for enzyme data and metabolic information. *Trends Biochem. Sci.* **27**, 54–56 (2002).
49. Kim, S. *et al.* PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* **47**, D1102–D1109 (2019).
50. Wicker, J. *et al.* enviPath – The environmental contaminant biotransformation pathway resource. *Nucleic Acids Res.* gkv1229 (2015). doi:10.1093/nar/gkv1229
51. Latino, D. A. R. S. *et al.* Eawag-Soil in enviPath: a new resource for exploring regulatory pesticide soil biodegradation pathways and half-life data. *Environ. Sci. Process. Impacts* **19**, 449–464 (2017).
52. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? *Nat. Biotechnol.* **28**, 245–8 (2010).
53. Henry, C. S., Broadbelt, L. J. & Hatzimanikatis, V. Thermodynamics-based metabolic flux analysis. *Biophys. J.* **92**, 1792–1805 (2007).
54. Salvy, P. *et al.* pyTFA and matTFA: a Python package and a Matlab toolbox for Thermodynamics-based Flux Analysis. *Bioinformatics* **35**, 167–169 (2018).
55. Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.* **7**, 20 (2015).
56. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **36**, D190–D195 (2007).
57. Rahman, S. A., Cuesta, S. M., Furnham, N., Holliday, G. L. & Thornton, J. M. EC-BLAST: a tool to automatically search and compare enzyme reactions. *Nat. Methods* **11**, 171–174 (2014).
58. Yamanishi, Y., Hattori, M., Kotera, M., Goto, S. & Kanehisa, M. E-zyme: predicting potential EC numbers from the chemical transformation pattern of substrate-product pairs. *Bioinformatics* **25**, i179–i186 (2009).
59. Liu, H. & Lu, T. Autonomous production of 1,4-butanediol via a de novo biosynthesis pathway in engineered *Escherichia coli*. *Metab. Eng.* **29**, 135–141 (2015).
60. Wang, J. *et al.* Rational engineering of diol dehydratase enables 1,4-butanediol biosynthesis from xylose. *Metab. Eng.* **40**, 148–156 (2017).
61. Kirby, J. & Keasling, J. D. Biosynthesis of Plant Isoprenoids: Perspectives for Microbial Engineering. *Annu. Rev. Plant Biol.* **60**, 335–355 (2009).
62. Peralta-Yahya, P. P. *et al.* Identification and microbial production of a terpene-based advanced biofuel. *Nat. Commun.* **2**, 483 (2011).

63. Dickschat, J. S., Brock, N. L., Citron, C. A. & Tudzynski, B. Biosynthesis of Sesquiterpenes by the Fungus *Fusarium verticillioides*. *ChemBioChem* **12**, 2088–2095 (2011).
64. Bohlmann, J., Crock, J., Jetter, R. & Croteau, R. Terpenoid-based defenses in conifers: cDNA cloning, characterization, and functional expression of wound-inducible (E)-alpha-bisabolene synthase from grand fir (*Abies grandis*). *Proc. Natl. Acad. Sci. U. S. A.* **95**, 6756–61 (1998).
65. Clomburg, J. M., Qian, S., Tan, Z., Cheong, S. & Gonzalez, R. The isoprenoid alcohol pathway, a synthetic route for isoprenoid biosynthesis. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 12810–12815 (2019).
66. Davies, F. K., Work, V. H., Beliaev, A. S. & Posewitz, M. C. Engineering Limonene and Bisabolene Production in Wild Type and a Glycogen-Deficient Mutant of *Synechococcus* sp. PCC 7002. *Front. Bioeng. Biotechnol.* **2**, 21 (2014).
67. Phelan, R. M., Sekurova, O. N., Keasling, J. D. & Zotchev, S. B. Engineering Terpene Biosynthesis in *Streptomyces* for Production of the Advanced Biofuel Precursor Bisabolene. *ACS Synth. Biol.* **4**, 393–399 (2015).
68. Orth, J. D. *et al.* A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism--2011. *Mol. Syst. Biol.* **7**, 535–535 (2014).
69. Mo, M. L., Palsson, B. Ø. & Herrgård, M. J. Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC Syst. Biol.* **3**, 37 (2009).
70. Broddrick, J. T. *et al.* Unique attributes of cyanobacterial metabolism revealed by improved genome-scale metabolic modeling and essential gene analysis. *Proc. Natl. Acad. Sci.* **113**, E8344–E8353 (2016).
71. Newman, D. J. & Cragg, G. M. Natural products as sources of new drugs over the 30 years from 1981 to 2010. *J. Nat. Prod.* **75**, 311–335 (2012).
72. Ro, D.-K. *et al.* Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* **440**, 940–943 (2006).
73. Galanie, S., Thodey, K., Trenchard, I. J., Interrante, M. F. & Smolke, C. D. Complete biosynthesis of opioids in yeast. *Science (80-.)*. **349**, 1095–1100 (2015).
74. Li, Y. *et al.* Complete biosynthesis of noscapine and halogenated alkaloids in yeast. *Proc. Natl. Acad. Sci.* **115**, E3922–E3931 (2018).
75. Lin, M., Chueh, F., Hsieh, M. & Chen, C. ANTIHYPERTENSIVE EFFECTS OF dl-TETRAHYDROPALMATINE: AN ACTIVE PRINCIPLE ISOLATED FROM CORYDALIS. *Clin. Exp. Pharmacol. Physiol.* **23**, 738–745 (1996).
76. Chung Leung, W., Zheng, H., Huen, M., Lun Law, S. & Xue, H. Anxiolytic-like action of orally administered dl-tetrahydropalmatine in elevated plus-maze. *Prog. Neuro-Psychopharmacology Biol. Psychiatry* **27**, 775–779 (2003).
77. Mantsch, J. R. *et al.* Levo-tetrahydropalmatine attenuates cocaine self-administration and cocaine-induced reinstatement in rats. *Psychopharmacology*

- (Berl). **192**, 581–591 (2007).
78. Bastian, M., Heymann, S. & Jacomy, M. Gephi: An Open Source Software for Exploring and Manipulating Networks. (2009).
 79. Carbonell, P., Gök, A., Shapira, P. & Faulon, J.-L. Mapping the patent landscape of synthetic biology for fine chemical production pathways. *Microb. Biotechnol.* **9**, 687–695 (2016).
 80. Lipinski, C. A. Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov. Today Technol.* **1**, 337–341 (2004).
 81. Kim, S., Thiessen, P. A., Cheng, T., Yu, B. & Bolton, E. E. An update on PUG-REST: RESTful interface for programmatic access to PubChem. *Nucleic Acids Res.* **46**, W563–W570 (2018).
 82. Sayers, E. A General Introduction to the E-utilities. (2010).
 83. Morishige, T., Dubouzet, E., Choi, K.-B., Yazaki, K. & Sato, F. Molecular cloning of columbamine O-methyltransferase from cultured *Coptis japonica* cells. *Eur. J. Biochem.* **269**, 5659–5667 (2002).
 84. Dang, T.-T. T. & Facchini, P. J. Characterization of three O-methyltransferases involved in noscapine biosynthesis in opium poppy. *Plant Physiol.* **159**, 618–31 (2012).
 85. Li, J. *et al.* Production of plant-specific flavones baicalein and scutellarein in an engineered *E. coli* from available phenylalanine and tyrosine. *Metab. Eng.* **52**, 124–133 (2019).
 86. Jones, J. A. *et al.* Experimental and computational optimization of an *Escherichia coli* co-culture for the efficient production of flavonoids. *Metab. Eng.* **35**, 55–63 (2016).
 87. Luo, X. *et al.* Complete biosynthesis of cannabinoids and their unnatural analogues in yeast. *Nature* **1** (2019). doi:10.1038/s41586-019-0978-9
 88. Zhang, C., Chen, X., Lindley, N. D. & Too, H.-P. A “plug-n-play” modular metabolic system for the production of apocarotenoids. *Biotechnol. Bioeng.* **115**, 174–183 (2018).
 89. Finley, S. D., Broadbelt, L. J. & Hatzimanikatis, V. Computational framework for predictive biodegradation. *Biotechnol. Bioeng.* **104**, 1086–1097 (2009).
 90. de Lorenzo, V. Systems biology approaches to bioremediation. *Curr. Opin. Biotechnol.* **19**, 579–589 (2008).
 91. Dvořák, P., Nikel, P. I., Damborský, J. & de Lorenzo, V. Bioremediation 3.0: Engineering pollutant-removing bacteria in the times of systemic biology. *Biotechnol. Adv.* **35**, 845–866 (2017).
 92. El Fantroussi, S. & Agathos, S. N. Is bioaugmentation a feasible strategy for pollutant removal and site remediation? *Current Opinion in Microbiology* **8**, 268–275 (2005).
 93. Liu, L., Bilal, M., Duan, X. & Iqbal, H. M. N. Mitigation of environmental pollution by

- genetically engineered bacteria — Current challenges and future perspectives. *Sci. Total Environ.* **667**, 444–454 (2019).
94. Eysers, L. *et al.* Environmental genomics: Exploring the unmined richness of microbes to degrade xenobiotics. *Applied Microbiology and Biotechnology* **66**, 123–130 (2004).
 95. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
 96. Atashgahi, S. *et al.* Prospects for harnessing biocide resistance for bioremediation and detoxification. *Science (80-.)*. **360**, 743–746 (2018).
 97. Hiraga, K., Taniguchi, I., Yoshida, S., Kimura, Y. & Oda, K. Biodegradation of waste PET: A sustainable solution for dealing with plastic pollution. *EMBO Rep.* e49365 (2019). doi:10.15252/embr.201949365
 98. Shim, H. & Yang, S. T. Biodegradation of benzene, toluene, ethylbenzene, and o-xylene by a coculture of *Pseudomonas putida* and *Pseudomonas fluorescens* immobilized in a fibrous-bed bioreactor. *J. Biotechnol.* **67**, 99–112 (1999).
 99. Rodrigues, D. F., Sakata, S. K., Comasseto, J. V., Bicego, M. C. & Pellizari, V. H. Diversity of hydrocarbon-degrading *Klebsiella* strains isolated from hydrocarbon-contaminated estuaries. *J. Appl. Microbiol.* **106**, 1304–1314 (2009).
 100. Das, S. *et al.* Characterization of Three *Mycobacterium* spp. with Potential Use in Bioremediation by Genome Sequencing and Comparative Genomics. *Genome Biol. Evol.* **7**, 1871–1886 (2015).
 101. Prijambada, I. D., Negoro, S., Yomo, T. & Urabe, I. Emergence of nylon oligomer degradation enzymes in *Pseudomonas aeruginosa* PAO through experimental evolution. *Appl. Environ. Microbiol.* **61**, 2020–2 (1995).
 102. Howard, G. T. & Blake, R. C. Growth of *Pseudomonas fluorescens* on a polyester-polyurethane and the purification and characterization of a polyurethanase-protease enzyme. *Int. Biodeterior. Biodegrad.* **42**, 213–220 (1998).
 103. Zhu, B. *et al.* Survival and chlorpyrifos-degradation of strain *Cupriavidus taiwanensis* Lux-X1 in different type soils. (2013).

Chapter 6 Conclusions and perspectives

The purpose of this final chapter is to summarize the presented achievements, and to discuss future perspectives regarding the development of computational methods to model, predict and mine metabolism. In particular, we will emphasize the opportunities provided by the presented methods to better characterize and design metabolism for metabolic engineering, drug discovery and synthetic biology applications in the future.

6.1 Conclusions

As a whole, this thesis explores different aspects and applications of the computational representation of enzymatic catalysis. The mechanistic knowledge of enzymatic action enabled us to draw an atom-level representation of metabolism, and thanks to the promiscuous nature of the encoded reaction mechanisms, we could explore the potential biochemical reaction space in different biological contexts. In the following paragraphs, we will recapitulate the major learnings from each aspect examined in the presented work.

To begin with, we employed the reaction rules representing enzymatic reaction mechanisms to map atoms throughout metabolic reactions, pathways and networks (Chapter 2). The atom-mapped networks could then be used to simulate ^{13}C tracer experiments in *E. coli*, by integrating carbon-labeled isotopomer networks into a constraint-based modeling approach. Although this novel framework will need further testing, the preliminary results are promising. Thanks to the systematic reduction and organization of the metabolic model used for *the in silico* labeling experiments, our approach was shown to be readily applicable to non-model organisms such as the malaria parasite *Plasmodium falciparum*.

The experience and insights gained from tracing atoms through metabolic reactions and networks fueled the development of a novel approach for pathway search (Chapter 3). The presented tool, named NICEpath, takes into account the conservation of atoms between a substrate and product to construct a searchable graph structure of a biochemical reaction network. We validated the proposed metric, called Conserved Atom Ration (CAR), through comparison with a manually curated database of substrate-product pairs obtained from the KEGG database. We further showed that NICEpath can be applied to reliably extract biologically relevant pathways from large metabolic networks such as KEGG or ATLAS.

Next, we explored the potential of the generalized enzymatic reaction rules, encoding the catalytic elasticity of enzymes, to predict novel, hypothetical biochemical reactions (Chapter 4). The resulting database series, named ATLASx, contains known and novel biochemical reactions within different compound scopes. The first ATLAS of Biochemistry database was built by predicting novel biochemical reactions between known KEGG compounds, relying on biochemical knowledge encoded in the reaction rules. As of today, more than 150 research groups have requested access to the ATLAS database. Furthermore, some of the predicted reactions have been validated experimentally, highlighting the interest of the scientific community in such a resource and its utility to advance metabolic engineering. The continued expansion of ATLAS towards all biological and bioactive compounds (bioATLAS), and further to chemicals (chemATLAS), has shown first, promising results by proposing the integration of more than 900,000 bioactive compounds and close to two million chemicals into the biochemical reaction space. To provide access to this wealth of data, we developed an interactive web interface that allows to search for compounds, reactions and pathways, the latter one enabled by a back-end architecture featuring the previously developed NICE-path method.

Finally, we applied the presented tools and methods to three different engineering and research problems. We first introduced a refined retrobiosynthesis workflow for industrial and scientific applications, and we then applied it to the commodity chemical 1,4-butanediol and to the biofuel bisabolene to predict potential biosynthesis pathways. We further computationally explored the potential derivatives of the noscapine pathway, we assessed their popularity and finally chose the pharmaceutical compound tetrahydropalmatine for bioproduction in yeast. In a last study, we adapted the retrobiosynthesis tools to evaluate the biodegradability of xenobiotics, for which we additionally developed a method to identify and rank organisms by their ability to degrade such compounds.

6.2 Future perspectives

The insights and learnings from this work lay the foundation for future developments and investigations, out of which two major aspects are discussed in the following: (i) The increased consideration of non-model organisms in computational approaches, and (ii) the improvement of the availability of computational pathway design tools for the scientific community.

6.2.1 Towards non-model organisms

In the presented work, we mainly discussed and applied the developed methods on a small selection of well-studied model organisms, such as *E. coli* and *S. cerevisiae*. The two exceptions are the application of iAM.NICE to study the malaria parasite *Plasmodium falciparum*, and our first steps in the territory of biodegradation where we screened the sequence database UniProtKB for organisms potentially capable of biodegradation. These two studies highlighted the value of our tools for the analysis of non-model organisms, as well as the importance to consider non-model organisms in our workflows. A first step towards this

objective is made in NICEpath, which can take a GEM as an input to find pathways connecting to metabolites present in the organism. Furthermore, systematic reduction approaches such as redGEM and lumpGEM enable the consistent organization of genome-scale models, which is particularly useful to analyze non-model organisms. While the availability of genome-scale models still suffers from a bias towards easily culturable model organisms, high-throughput sequencing techniques have helped to populate sequence databases such as UniProtKB with genomic sequences obtained from the environment. The available sequence data can be used to integrate non-culturable, non-model organisms into our predictions and considerations. As a conclusion, broadening our horizon towards non-model organisms will hopefully lead to new discoveries in the study of microbial communities, vector-borne diseases, and bioengineering of non-standard organisms such as extremophiles for bioengineering in salt water conditions.

6.2.2 Next-generation pathway design

Designing metabolic pathways for biosynthesis in an accurate and reliable manner is challenging. In an ideal case, all the necessary methods would be integrated in one single tool that can predict biosynthesis pathways, evaluate them in the context of a host organisms, find adapted enzymes and provide codon-optimized gene sequences that can be readily used to transform the host organisms for the bioproduction of a compound of interest. This idea has been discussed by Nielsen and Keasling, who imagined a biological Computer-Aided Design (CAD) tool, as it is widely used in other engineering disciplines such as mechanical, electrical and civil engineering, for metabolic engineering¹. The output of such a BioCAD would be an experimental recipe that enables bioengineers to create new strains for bioproduction of any desired chemical compound. Even though such a tool is not yet reality for metabolic engineering, it has been achieved for chemical synthesis. The Chematica platform proposes a computer-assisted planning of synthesis routes based on the concepts of retrosynthesis and chemical reaction rules². Chematica has been successfully applied to generate synthesis recipes that could be directly used by chemists to perform complex syntheses of a range of benchmark chemicals, without requiring any prior experience in multi-step organic synthesis³. These predictions were generated within 15-20 minutes thanks to the integration of several mathematical and computational techniques such as graph theory, linear programming, artificial intelligence and expert-curated chemical knowledge. A tool providing such accurate predictions in a short time would be desirable for metabolic engineering, and it would help to accelerate the expansion of range of biosynthetically produced chemicals.

To achieve a comparable success rate for biological systems however, we first need to overcome multiple challenges, most of them due to the complexity of biological systems. Currently, available computational techniques were not sufficient to design *de novo* biosynthetic pathways in a fast and reliable way comparable to Chematica. For example, knowledge gaps in biology affect the accuracy of biosynthetic predictions, and additional factors should be considered such as choice of the best chassis organism, uncertainties in the activity of enzymes (*e.g.*, kinetic properties, substrate promiscuity), and metabolic

network properties. To consider these factors, an intelligent integration of available knowledge is required to predict biosynthesis pathways in a way comparable to chemical synthesis. We estimate that by efficiently integrating available biochemical data and by employing well-calibrated prediction tools for reactions, enzymes, network structures and regulatory mechanisms, we can one day achieve the same efficiency in predicting biosynthetic routes in metabolic engineering as it is now possible for chemical synthesis. Future work will also help us to automatically detect and, if possible, fill knowledge gaps in metabolism, such as missing biosynthesis steps in secondary metabolism, underground enzymatic activity and metabolic dark matter, on a large scale.

We believe that databases like ATLAS can help us move towards the objective BioCAD. Like Chematica, the ATLASx databases are built on a solid basis of chemical knowledge incorporated by generalized reaction rules. ATLASx provides hypothetical reactions, annotated with putative enzymes, and is searchable thanks to an integrated, efficient pathway search. In the future, we will have to find systematic ways to reliably integrate novel biochemical compounds, and we will further invest efforts into making our databases more user-friendly, improve the visualization of pathways and make it more accessible for non-specialists in general. Providing additional analysis tools on our site, such as thermodynamic feasibility calculations in a chassis organism of choice and integrated tools for extended enzyme prediction, will additionally improve the value of ATLASx and bring the idea of a BioCAD one step closer to reality.

References

1. Nielsen, J. & Keasling, J. D. Engineering Cellular Metabolism. *Cell* **164**, 1185–1197 (2016).
2. Szymkuć, S. *et al.* Computer-Assisted Synthetic Planning: The End of the Beginning. *Angew. Chemie Int. Ed.* **55**, 5904–5937 (2016).
3. Klucznik, T. *et al.* Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory. *Chem* **4**, 522–532 (2018).

Appendix

Table A1: Lumped reactions for 20 BBBs and their reconstruction in iAM.NICE

[See Appendix.xlsx – A1]

Table A2: Reactant pairs in KEGG with non-zero CAR and corresponding RPAIR annotation

[See Appendix.xlsx – A2]

Table A3: Noscapine pathway derivatives ranked by popularity

[See Appendix.xlsx – A3]

The appendix is published on the Zenodo platform and accessible at <https://zenodo.org/record/3703123> (DOI: 10.5281/zenodo.3703123).

Tools and databases are hosted on <https://lcsb-databases.epfl.ch/>.

Curriculum Vitae

CONTACT AND PERSONAL INFORMATION

Jasmin Hafner Address: Route Aloys-Fauquez 109, CH-1018 Lausanne, Switzerland

E-mail: jasmin.hafner@epfl.ch Tel.: [+41 21 693 76 43](tel:+41216937643)

Date of birth: 19 May 1991 Place of birth: Lenzburg, AG, Switzerland Nationality: Swiss

EMPLOYMENT HISTORY

Aug. 2010 and Aug. 2011	Research internship with focus on biodiversity at the Swiss center for agricultural research Agroscope in Reckenholz-Tänikon, Switzerland (2 times, one month each)
March 2013 - Jan. 2015	Part time customer consultant, secretary and translator (French-German) at the start-up company Yoocos Sàrl, Renens, Switzerland
July 2013 – July 2015	Part time job as an accompanying instructor for mentally handicapped children and adults at Solidarité-Handicap mental, Lausanne, Switzerland
Aug. 2015 – Feb. 2020	PhD candidate at the School of Chemistry and Chemical Engineering (EDCH), École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

EDUCATION

1997 - 2006	Primary and secondary school in Balsthal, SO, Switzerland,
2006 - 2010	Matura at Kantonsschule Solothurn, Switzerland
1010 - 2013	B.Sc. in Biology, University of Lausanne, Switzerland
Aug. 2012 – Dec. 2012	Student exchange including scholarship to Arizona State University (ASU), Tempe AZ, U.S.A.
2013 - 2015	M.Sc. in Molecular Biology - mention Bioinformatics, University of Lausanne, Switzerland Master Thesis: “ <i>In silico</i> reconstruction of atom-level metabolic networks” at LCSB, EPFL. Advisor: Prof. Vassily Hatzimanikatis

RESEARCH AND TRAINING

- | | |
|---------------------|--|
| Nov 2016 – Feb 2017 | Research exchange at Stanford University, Stanford, CA, U.S.A. Collaboration with Prof. Christina Smolke |
| Aug 2017 – Feb 2018 | Industrial collaboration with l'Oréal Advanced Research, Aulnay-Sous-Bois, France |
| Jan 2018 – Dec 2018 | Industrial collaboration and internship at Nestlé Research Center (NRC), Lausanne, Switzerland |
-

PROFESSIONAL QUALIFICATIONS

- Languages: German (mother tongue), English (fluent), French (fluent)
 - Computational Skills: Programming (Python, C++), Databases (SQL), Data analysis (MATLAB, R), Web development (basics in JavaScript, PHP, HTML)
-

GRANTS AND AWARDS

- Student/Young Investigator Poster Award, Category Runner-Up, Metabolic Engineering 11, 2016, Awaji City, Japan, sponsored by Wiley and Biotechnology Journal
 - All SystemsX.ch Day 2016, Best Poster Award – 1st place, category PhD student, Bern, Switzerland
 - Research exchange grant for three months, EPFL-Stanford Program, 2016, sponsored by Firmenich
-

PUBLICATIONS

- Hadadi, N., **Hafner, J.**, Shajkofci, A., Zisaki, A. & Hatzimanikatis, V. ATLAS of Biochemistry: A Repository of All Possible Biochemical Reactions for Synthetic Biology and Metabolic Engineering Studies. *ACS Synth. Biol.* **5**, 1155–1166 (2016).
 - Hadadi, N., **Hafner, J.**, Soh, K. C. & Hatzimanikatis, V. Reconstruction of biological pathways and metabolic networks from in silico labeled metabolites. *Biotechnol. J.* **12**, 1600464 (2017).
 - **Hafner, J.**, Mohammadi-Peyhani, H., Svenshnikova, A., Scheidegger, A. & Hatzimanikatis, V. Updated ATLAS of biochemistry with new metabolites and improved enzyme prediction power. *Accepted by ACS Synthetic biology* (2020)
-

PERSONAL INTERESTS

Rock climbing, hiking, traveling, literature