Network Design and Analysis for Multi-Enzyme Biocatalysis

Dissertation
zur Erlangung des Grades
der Doktorin der Naturwissenschaften
der Naturwissenschaftlich-Technischen Fakultäten
der Universität des Saarlandes

von

Lisa Katharina Schuh

Saarbrücken
2020

Tag des Kolloquiums:	21.10.2020
Dekan:	Prof. Dr. Guido Kickelbick
Berichterstatter:	Prof. Dr. Elmar Heinzle
	Prof. Dr. Volkhard Helms
Vorsitzender:	Prof. Dr. Uli Müller
Akad. Mitarbeiter:	Dr.-Ing. Michael Kohlstedt

七転八起

Zusammenfassung

In vitro Synthese ist eine biotechnologische Alternative zu klassischen chemischen Katalysen. Der manuelle Entwurf von mehrstufigen Biosynthesewegen ist jedoch sehr anspruchsvoll, vor allem wenn Enzyme verschiedener Organismen beteiligt sind. Daher besteht ein Bedarf an Methoden, die helfen solche Synthesewege *in silico* zu entwerfen und die in der Lage sind große Mengen biologischer Daten zu bewältigen - insbesondere in Hinblick auf die Rekonstruktion genomskaliger metabolischer Netzwerkmodelle und die Pfadsuche in solchen Netzwerken.

In dieser Arbeit wird ein Algorithmus zur Pfadsuche zu einem Zielprodukt ausgehend von beliebigen Substraten präsentiert. Der Algorithmus basiert auf einem gemischt-ganzzahligen linearen Programm, das Graphtopologie mit Reaktionsstöchiometrien kombiniert. Die Pfadkandidaten werden anhand verschiedener Kriterien geordnet, um die am besten geeigneten Kandidaten für die Synthese zu finden. Außerdem wird ein umfassender Workflow für die Rekonstruktion metabolischer Netzwerke basierend auf der Datenbank KEGG sowie thermodynamischen Daten vorgestellt. Dieser umfasst einen Filter, der anhand verschiedener Kriterien geeignete Reaktionen auswählt. Der Workflow wird zum Erstellen einer organismusübergreifenden Netzwerkrekonstruktion, sowie Netzwerken einzelner Organismen genutzt. Diese Modelle werden mit graphentheoretischen Methoden analysiert. Es wird diskutiert, wie die Ergebnisse für die Planung von biosynthetischen Produktionswegen genutzt werden können.

Abstract

In vitro synthesis is a biotechnological alternative to classic chemical catalysts. However, the manual design of multi-step biosynthesis routes is very challenging, especially when enzymes from different organisms are involved. There is therefore a demand for *in silico* tools to guide the design of such synthesis routes using computational methods for the path-finding, as well as the reconstruction of suitable genome-scale metabolic networks that are able to harness the growing amount of biological data available.

This work presents an algorithm for finding pathways from arbitrary metabolites to a target product of interest. The algorithm is based on a mixed-integer linear program (MILP) and combines graph topology and reaction stoichiometry. The pathway candidates are ranked using different ranking criteria to help finding the best suited synthesis pathway candidates. Additionally, a comprehensive workflow for the reconstruction of metabolic networks based on data of the Kyoto Encyclopedia of Genes and Genomes (KEGG) combined with thermodynamic data for the determination of reaction directions is presented. The workflow comprises a filtering scheme to remove unsuitable data. With this workflow, a pan-organism network reconstruction as well as single organism network models are established. These models are analyzed with graph-theoretical methods. It is also discussed how the results can be used for the planning of biosynthetic production pathways.

Contents

Part I

Introduction and Background

CHAPTER 1

Multi-Enzyme Biocatalysis

The following chapter is based on the published review article

> HEINZLE, E., WEYLER, C., KRAUSER, S., & BLASS, L. K. (2013): Directed multistep biocatalysis using tailored permeabilized cells. A.-P. ZENG (Ed.), *Advances in biochemical engineering/biotechnology* (pp. 185–234). Springer Berlin Heidelberg. https://doi.org/10.1007/10_2013_240

that I co-authored.

In billions of evolutionary steps, nature developed an impressive set of strategies to create molecules with a wide range of structures. Nearly every carbon-, nitrogen-, oxygen-, or sulfur-containing skeleton and functional group can be assembled in principle by bioconversions. A large number of different enzymatically catalyzed reactions support cellular growth and survival (LOPEZ-GALLEGO et al., 2010). Not only the substrate and reaction specificity but also the efficiency of enzymatic reactions are usually far beyond man-made chemical processes. Recent developments in biochemical research not only support a detailed mechanistic understanding of relationship of structure and reactivity, but they also allow extended targeted redesign and modification of enzymes. Even completely new functionalities can be designed and created with modern molecular and modeling tools (e.g. Diels-Alder synthesis with a *de novo* designed enzyme (SIEGEL et al., 2010)). The development and present status concerning biocatalysis - mostly based on engineered single enzymes - have been reviewed thoroughly elsewhere (BORNSCHEUER et al., 2012). Recent developments in the field of metabolic network research, both experimentally as well as computationally, open up new potentials for multi-step biocatalysis both *in vivo* as well as *in vitro*.

Presently, slightly more than 100 commercial applications use enzymes in industrial-scale processes (LIESE et al., 2000). Due to the usually high price, the time required for improving enzymes genetically, the often shorter development times required for organic chemistry alternatives, and the still widespread ignorance of biocatalysis in the field of organic chemistry, bioconversion processes are often not considered (WOHLGEMUTH, 2011). There is, however, a trend towards biotechnological processes as the ecological impact (E-factor) of industrial productions is gaining weight and public pressure demands a sustainable industry (AEHLE, 2004; DRAUZ et al., 2012; HEINZLE et al., 2006; WOHLGEMUTH, 2010). Biocatalytic processes often have a very low ecological impact, such as with selective oxidation of carbohydrates (SCHNEIDER et al., 2012), but in some cases chemical alternatives are similar or even better (KUHN et al., 2010).

Although *in vivo* synthesis using whole, viable microorganisms provides complex products from simple and cheap raw materials by fermentation, it is limited by the fitness and tolerance of the organism and by cellular transport processes. Modern metabolic engineering methods provide a whole toolbox comprising computational and molecular tools for directed design and optimization of production pathways. In most cases, these allow the conversion of a poorly producing native organism into a highly efficient producer strain. However, transport barriers, bottlenecks in the metabolism, toxic side effects, and the usually required complex downstream processing of resulting mixtures of product and growth medium limit the industrial applications (Figure 1.1, Case C).

In vitro synthesis, on the other hand, serves as a biotechnological alternative to the classic chemical catalysts. Engineered for the highest activity, stability, and substrate spectrum, enzymes provide the highest turnover rates, simultaneously working with outstanding selectivity (BORNSCHEUER et al., 2012). However, in case of complex syntheses, the use of enzymes is restricted by required optimal conditions for each enzyme and potential intermediate clean-up or buffer change between individual steps (Figure 1.1, Case A). Additionally, the regeneration of cofactors, such as nicotinamide adenine dinucleotide (phosphate) (NAD(P)H), limits this type of application.

A further alternative is the one-pot synthesis using multiple enzymes; however, they require extended optimization of enzymes to operate at the same pH and buffer concentrations (Figure 1.1, Case B). This approach can also be taken using cell hydrolysates of suitable strains, as has been reviewed elsewhere (YOU et al., 2013). On the other hand, various approaches use synthetic assemblies of enzymes, such as in emulsions, using scaffolds,

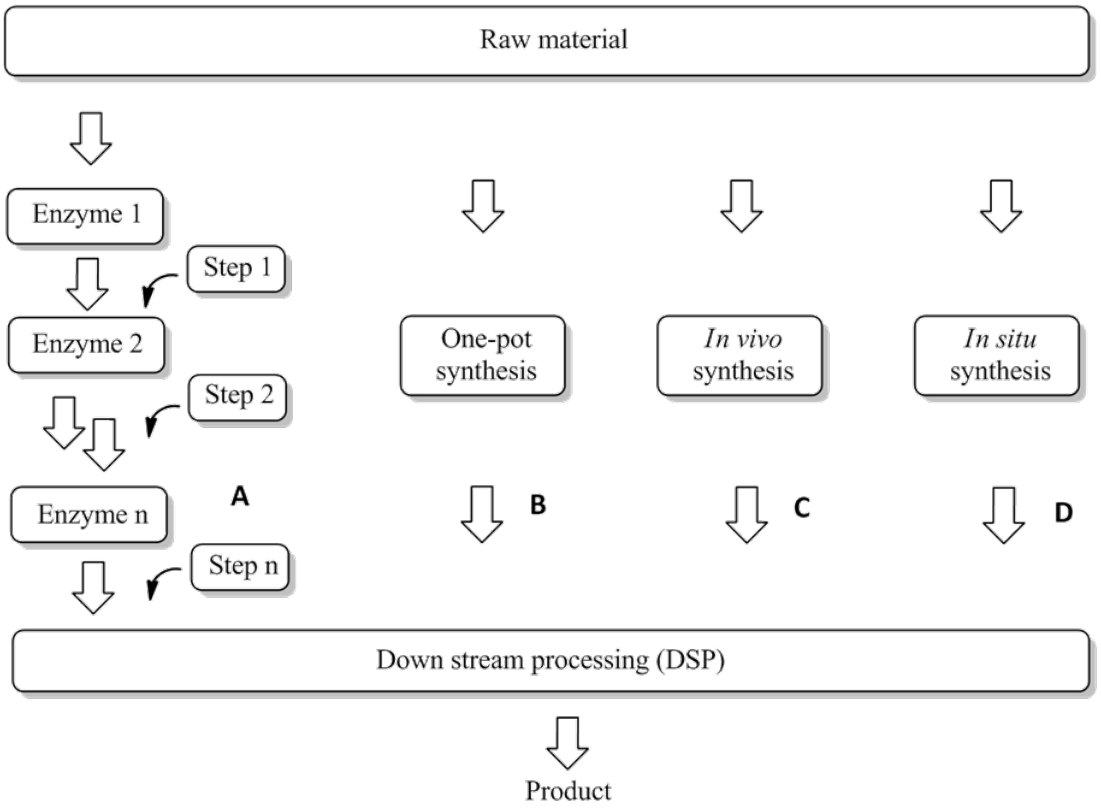


Figure 1.1: Types of multi-step biosynthetic processes. A: Synthesis using multiple enzymes in separate processes, B: Synthesis with all enzymes reacting in one-pot, C: *In vivo* synthesis using living cells in fermentation processes, D: *In situ* synthesis using permeabilized cells.

tethering to surfaces, or covalent binding to achieve one-pot synthesis biocatalysis (MOSES et al., 2013).

Yet another strategy uses permeabilized cells - often called *in situ* synthesis (Figure 1.1, Case D). Permeabilized cell membranes allow diffusion of small compounds between the intracellular space and the surrounding reaction buffer while large biopolymers (i.e. proteins and deoxyribonucleic acid (DNA)) remain trapped inside the microenvironment of the cell. Contrary to using cell hydrolysates, which has a very long tradition (YOU et al., 2013), optimal permeabilization will keep the enzymes in their native macromolecular environment and does not cause any denaturation of enzymes by the permeabilizing agent. In this way, the macromolecular crowding effects that are expected to modify protein activities (MINTON, 2006), such as channeling (MONTI et al., 2009), are preserved in their original status. Removing all small metabolites and cofactors represents a kind of reset of the

metabolic network, permitting directed conversions by the selection of appropriate substrate combinations. In general, this can be combined with careful tailoring of the enzymatic outfit of a cell, thus increasing selectivity of bioconversion using permeabilized cells. Network changes may involve gene deletions, gene amplification, or heterologous gene expression. Additionally, selective inhibitors might be used to block undesired side reactions (KRAUSER et al., 2012). An interesting alternative concept (YE et al., 2012) uses enzymes from thermophilic organisms that are expressed in a mesophilic organism. Cells are cultivated and then heated to rupture the cells and inactivate enzymes that are not desired for the *in vitro* biocatalytic conversion.

The term '*in situ* synthesis' was introduced in the early 1960s, indicating that macromolecules remain in their original macromolecular environment. Prokaryotic and eukaryotic cells can be permeabilized, but the permeabilization procedure depends on the composition of the cell wall and has to be optimized for each cell type. Early studies on permeabilized cells by Felix showed promising results. Felix concluded that permeabilized cells can be produced quickly and simply and can be used several times, thus requiring less energy for the synthesis of biomass (FELIX, 1982). These studies on synthesis with permeabilized cells never achieved appropriate acknowledgment, however - likely because of the missing genetic and metabolic engineering tools at that time. The available tools have dramatically changed since then, and it seems obvious that synthesis with permeabilized cells will provide an alternative method, thus closing the gap between *in vivo* and *in vitro* biosynthesis.

The enormous increase in DNA sequencing power has recently created an overwhelming wealth of genome and metabolic network information of a large number of single (micro)organisms but also of microbial habitats using metagenome analysis. In parallel, computational tools for handling and exploring this vast amount of data have been developed at a high rate. However, detailed biochemical knowledge of enzyme characteristics is lagging far behind. Nevertheless, genome and enzyme databases provide an enormous amount of data that may be explored for permeabilized cell synthesis. Whole genome metabolic networks become increasingly available - a few of them already carefully curated. Metabolic regulation is also increasingly explored, but it requires considerably higher effort compared to sequencing. For some microorganisms, such as *Escherichia coli* and *Saccharomyces cerevisiae*, metabolic and regulatory networks are already fairly well understood, but we are still quite far away from the comprehensive understanding required for creating fully predictive models. This is even more the case or the majority of microorganisms. The metabolism of microorganisms may differ considerably. Nevertheless, they all share large parts of their central metabolism, particularly the 12 small precursor molecules representing the bottleneck of the bow-tie-shaped structure of metabolic networks (MA et al., 2003). These precursor molecules

serve as starting materials for all building blocks and polymers that can be synthesized in the metabolic network (Figure 1.2). Biopolymers constitute the major fraction of cellular

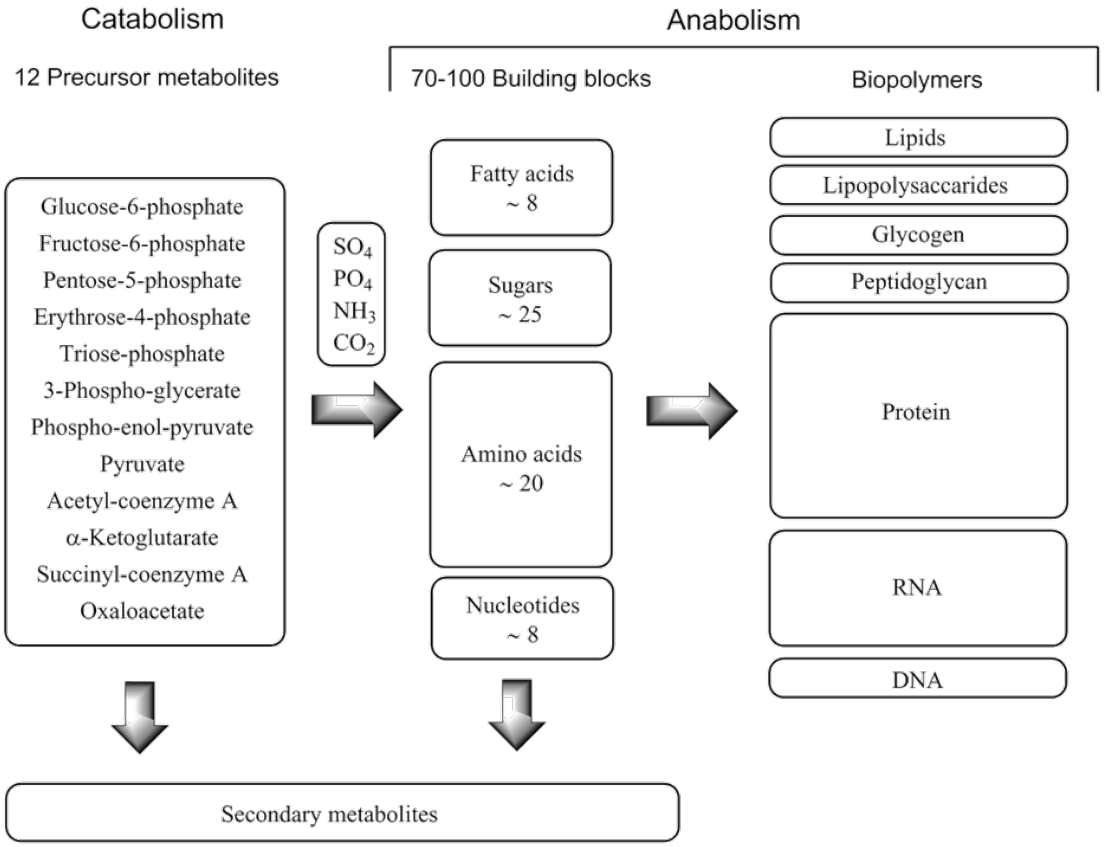


Figure 1.2: Overall structure of metabolism

biomass. Secondary metabolites are of great interest as pharmaceutically active compounds or precursors thereof. These are synthesized starting from precursor metabolites and building blocks. Up to now, this has in most cases been done with fully intact genetically engineered cells.

With the present knowledge, molecular and computational methods, and the advent of new possibilities of designing and engineering enzymes and whole metabolic pathways, a large field of applications opens up. Together with the long-known technique of permeabilization of cellular barriers (i.e. cell membranes and cell walls), new and intriguing opportunities for designing tailored biocatalyst and bioprocesses become accessible. Once such a biocatalyst is established, it can be produced easily by simple cultivation followed by permeabilization. Downstream processing would simply start with the removal of the biocatalyst, such as by

centrifugation. There are, however, major hurdles to be overcome as far as more complex biosynthesis is concerned. The most important are the supply of precursors and cosubstrates, such as adenosine 5'-triphosphate (ATP) or NAD(P)H; the achievement of selective biocatalysts, meaning the elimination of the manifold possible undesirable side reactions; and the intensification of processes to obtain high final product concentrations. To reach these goals, it is important to understand (i) the permeabilization process on a molecular basis, (ii) biosynthetic pathways and their regulation, (iii) supply of precursor molecules, (iv) regeneration of cosubstrates, (v) design and selection of enzymes as part of the biosynthetic processes, and (vi) design of biocatalytic pathways on the basis of the vast genomic and biochemical knowledge.

CHAPTER 2

Network Design

The following chapter is based on the published review article

> HEINZLE, E., WEYLER, C., KRAUSER, S., & BLASS, L. K. (2013): Directed multistep biocatalysis using tailored permeabilized cells. A.-P. ZENG (Ed.), *Advances in biochemical engineering/biotechnology* (pp. 185–234). Springer Berlin Heidelberg. https://doi.org/10.1007/10_2013_240

that I co-authored.

Directed, selective biosynthesis using either cell hydrolysates or permeabilized cells can be a straightforward process for shorter paths or small networks. It is, however, becoming increasingly challenging for longer biosynthetic paths with more compounds involved. The number of potentially undesired side reactions increases dramatically. Therefore, there is a great need to guide the design of such complex biocatalysts by using adequate computational tools and the rapidly increasing information available on mostly public databases. Panke and Bujara proposed an *in silico* tool for network topology analysis based on genome-scale metabolic network models to be applied for *in vitro* biocatalysis in cell-free systems (BUJARA et al., 2012). Starting out from the whole-genome scale metabolic reconstruction of *E. coli* (FEIST et al., 2007), they introduced several changes, particularly concerning transport and other membrane processes. Considering basic thermodynamic data and expression data from *E. coli*, they arrived at a model that could eventually predict interfering pathways for the production of dihydroxyacetone phosphate starting from glucose. The presently available examples of pathway prediction for biosynthesis using either cell extracts or permeabilized

cells are still very limited, but we expect that there will be a rapid increase of such studies in the very near future. There is, however, a whole series of studies available for living cells that are separated from the environment by their envelopes, providing selective transport of molecules in and out of the cells.

Modern planning and development of biochemical syntheses or novel synthesis routes in living organisms is effectively supported by the use of appropriate *in silico* tools. Such tools are increasingly available for pathway design in microorganisms and allow quick and directed engineering of living cells (KROEMER et al., 2006; NEUNER et al., 2011). These tools rely heavily on the existence and quality of the numerous biological databases containing information on different aspects such as genome sequences, enzyme data, or even whole pathways (Tables 2.1 to 2.5). Together with data from primary literature and further sources, this information can be used for the composition of network reconstructions of the organism of interest. Such networks can then be conveniently analyzed and developed further with different bioinformatic tools (Table 2.7 later in Section 2.2 Bioinformatic Tools). In particular, they can be used to design pathways or biosynthetic subnetworks useful for biocatalytic purposes, such as the *in situ* synthesis of a desired primary or secondary metabolite. With an increasing number of steps and increasing numbers of metabolites and coenzymes, the involved design becomes an increasingly complex task.

2.1 Databases

Biological databases can be classified into several different categories, such as biochemical databases, genome databases, protein or enzyme databases, pathway databases, or model databases (Tables 2.1 to 2.5). This classification is based on the biological content of the respective databases. However, an overlap of information can occur. For example, genome databases (Table 2.2) also contain protein sequence information.

Biochemical Databases

Table 2.1 lists different biochemical databases. *Rhea* (MORGAT et al., 2016) is a manually annotated, expert-curated reaction database with a main focus on enzyme-catalyzed reactions. It also contains other types of reactions. All reaction participants are linked to *Chemical Entities of Biological Interest (ChEBI)* (HASTINGS et al., 2015), which provides data such as structure, formula, and charge. All reactions in the database are stoichiometrically and charge-balanced and reaction directionality is added if it is available.

Table 2.1: Biochemical databases

Database	URL	Content
PubChem (S. Kim et al., 2019)	http://pubchem.ncbi.nlm.nih.gov	Chemical molecules and their activities against biological assays
ChEBI (Hastings et al., 2015)	http://www.ebi.ac.uk/chebi	Chemical Entities Of Biological Interest
TCDB (Saier Jr et al., 2015)	http://www.tcdb.org	Transporter Classification Database
Transport DB (Elbourne et al., 2016)	http://www.membranetransport.org	Transporter protein analysis database
SABIO-RK (Wittig et al., 2012)	http://sabio.villa-bosch.de	Biochemical reaction kinetics
Rhea (Morgat et al., 2016)	http://www.ebi.ac.uk/rhea	Manually annotated database of chemical reactions
MINE (Jeffryes et al., 2015)	https://minedatabase.mcs.anl.gov	Metabolic In Silico Network Expansion Databases

SABIO-RK (WITTIG et al., 2012), the biochemical reaction kinetics database, is a curated database containing biochemical reactions and their corresponding kinetics. It describes the participants and modifiers of the reactions as well as measured kinetic data, such as kinetic rate equations, embedded in an experimental and environmental context.

The *Transporter Classification Database (TCDB)* (SAIER JR et al., 2015) provides a functional and phylogenetic classification of membrane transport proteins. The classification system used is the transporter classification (TC) system that is analogous to the Enzyme Commission (EC) number for enzymes. The database is curated with data from over 15,000 published references. It contains over 18,000 unique protein sequences that are classified in more than 1,600 transporter families.

TransportDB (ELBOURNE et al., 2016) contains the predicted cell membrane transport protein complement for over 2760 organisms (bacteria, archaea, and eukaryota). The protein classification is done according to the TC classification system.

MINE (JEFFRYES et al., 2015) is a database containing predicted molecules that are likely to occur in reactions based on known metabolites and common biochemical reactions. The

prediction utilizes BNICE (HATZIMANIKATIS et al., 2005) and expert-curated reaction rules. The database contains more than 571,000 compounds.

ATLAS of Biochemistry (HADADI et al., 2016) is a database containing all possible theoretical biochemical reactions (more than 137,416 known and novel reactions) predicted with BNICE (HATZIMANIKATIS et al., 2005) using the means of enzyme reaction rules as well as other cheminformatic tools.

Genome Databases

Genome databases (Table 2.2) contain nucleotide sequences and functional annotations. *GenBank* (BENSON et al., 2013), run by the National Center for Biotechnology Information (NCBI), is a genetic sequence database of all publicly available DNA sequences. It contains the bibliographic and biological annotated sequences from almost 260,000 organisms. *NCBI Entrez Gene* (MAGLOTT et al., 2011) is a gene database containing a large variety of information that focuses on completely sequenced genomes. *NCBI Entrez Genome* (SAYERS et al., 2012) contains sequence and map data of more than 1,000 species or strains. The *Gene Ontology (GO)* (ASHBURNER et al., 2000; T. G. O. CONSORTIUM, 2019) focuses on the function genes.

Table 2.2: Genome databases

Database	URL	Content
GenBank® (Benson et al., 2013)	`http://www.ncbi.nlm.nih.gov/Genbank`	Annotated collection of all publicly available DNA sequences
NCBI Entrez Genome (Sayers et al., 2012)	`http://www.ncbi.nlm.nih.gov/sites/genome`	Sequence and map data from whole genomes of over 1,000 species and strains
NCBI Entrez Gene (Maglott et al., 2011)	`http://www.ncbi.nlm.nih.gov/gene`	Database of genes
GO (Ashburner et al., 2000; T. G. O. Consortium, 2019)	`http://www.geneontology.org`	The Gene Ontology
ENA (Hussein et al., 2018)	`http://www.ebi.ac.uk/ena`	European Nucleotide Archive

Protein and Enzyme Databases

Protein and enzyme databases (Table 2.3) collect functional information from proteins and enzymes. *Braunschweig Enzyme Database (BRENDA)* (JESKE et al., 2018), is a collection of

Table 2.3: Protein and enzyme databases

Database	URL	Content
BRENDA (Jeske et al., 2018)	http://www.brenda-enzymes.info	Comprehensive enzyme information System
Expasy-ENZYME (Bairoch, 2000)	http://www.expasy.org/enzyme	Enzyme nomenclature database
UniProt (U. Consortium, 2018)	http://www.uniprot.org	The Universal Protein Resource
PSORTdb (Peabody et al., 2016; Yu et al., 2011)	http://db.psort.org	Protein subcellular localizations for bacteria and archaea
ProLinks (Bowers et al., 2004)	http://prl.mbi.ucla.edu/prlbeta	Inferring functional linkages between proteins
STRING (Szklarczyk et al., 2019)	http://string-db.org	Search Tool for the Retrieval of Interacting Genes/Proteins
IntAct (Orchard et al., 2014)	http://www.ebi.ac.uk/intact	Molecular interaction database

functional and property data of enzymes. The majority of the contained data is manually extracted from primary literature and covers information in over 50 data fields, such as classification and nomenclature; reaction and specificity; information on function, structure, occurrence, preparation, and application of enzymes; and properties of mutants and engineered variants. Enzymes in BRENDA are linked to their respective pathways, source organism, and protein sequence, if deposited.

UniProt (U. CONSORTIUM, 2018), the universal protein resource, contains information on protein sequences and annotation data. It comprises four databases, namely the UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters (UniRef), the UniProt Archive (UniParc), and the UniProtMetagenomic and Environmental Sequences (UniMES) database on metagenomic and environmental data. UniProtKB is a collection of functional information on proteins together with annotation. The core data available for each protein are its amino acid sequence, protein name or description, taxonomic data, and citation information. Additionally, it contains as much annotation information as possible, such as ontologies,

classifications, and cross-references, together with an indication of annotation quality. The database consists of two sections: UniProtKB/SwissProt contains reviewed and manually annotated records, whereas UniProtKB/TrEMBL has data records that are unreviewed and automatically annotated and still await full manual annotation. In the April 2020 release, UniProtKB/Swiss-Prot contained more than 562,000 sequence entries and UniProtKB/TrEMBL contained more than 180,690,000 sequence entries. UniRef is a database providing clustered sets of sequences from UniProtKB, including splice variants and isoforms and selected UniParc records. Its purpose is to obtain the complete coverage of sequence space at several resolutions. UniParc contains most of the publicly available protein sequences.

The planning of a biochemical synthesis involves, besides other aspects, the determination of possible side reactions. Those unwanted reactions may lead to a decrease in the yield of the desired product and complicate the downstream processing. It would thus be favorable to find information on all reactions catalyzed by the enzyme of interest including thermodynamic and kinetic parameters. Some of the enzyme resources presented in Table 2.3 contain information on single enzymes. However, as of 2020, there is, to our knowledge, no database that presents such information in a systematic manner.

Model Databases

Model databases (Table 2.4) are repositories of mathematical models of biological systems. They contain models ranging from reconstructions of individual pathways up to genome-scale metabolic networks of organisms.

Table 2.4: Model databases

Database	URL	Content
BiGG (King, Lu, et al., 2015)	`http://bigg.ucsd.edu`	Knowledgebase of Biochemically, Genetically, and Genomically Structured Genome-Scale Metabolic Network Reconstructions
BioModels (Chelliah et al., 2014; Glont et al., 2017)	`https://www.ebi.ac.uk/biomodels`	Annotated Published Models
EcoCyc (Keseler et al., 2017)	`http://ecocyc.org`	*E. coli* K-12 MG1655
SGD (Cherry et al., 2011)	`http://www.yeastgenome.org`	*S. cerevisiae* Genome Database

BiGG Models is a knowledge base of biochemically, genetically, and genomically structured genome-scale metabolic network reconstructions (KING, LU, et al., 2015). As of Mai 2020, it contains 108 different genome-scale reconstructions from different organisms and share a standard nomenclature. BiGG allows browsing the model contents, the visualization of metabolic pathway maps with the Escher pathway visualization library (KING, DRÄGER, et al., 2015), as well as the export of all models into different standard formats.

BioModels (CHELLIAH et al., 2014; GLONT et al., 2017) is a repository for computational models of biological systems. In August 2019, it contained more than 10,000 manually curated, auto-generated or non-curated models. The database features browsing of models through lists, based on GO terms or using the taxonomy annotations as well as model search. The model presentation gives access to all information stored about a model. All models can be exported in various file formats or represented graphically. Basic model simulation is also possible. The functionality of BioModels can also be accessed by other software tools through its web services.

Pathway Databases

Pathway databases (Table 2.5) contain data on biochemical pathways, their reactions, and components that are involved in them and their corresponding interactions, thus describing the biochemistry of metabolic processes. These databases offer the possibility of providing several types of information in the context of graphical representation of the pathways in pathway maps.

BioCyc (KARP et al., 2017) is a collection of more than 17,000 pathway/genome databases in version 23.5. Each database contains the genome and metabolic pathways of a single organism. Based on the quality of the data, the databases are divided into three tiers. Tier 1 contains databases that are curated based on literature data. Tier 2 and tier 3 databases contain computationally predicted metabolic pathways, predictions as to which genes code for missing enzymes in metabolic pathways, and predicted operons. Tier 2 undergoes moderate curation and tier 3 is not curated at all.

The *KEGG* (KANEHISA et al., 2012; KANEHISA et al., 2018) is a curated database resource that integrates genomic, chemical, and systemic function information of various organisms. Its knowledge base consists of 18 main databases in the 5 categories systems information (pathway, brite, module); genomic information (orthology, genome, genes, ssdb); chemical

Table 2.5: Pathway databases

Database	URL	Content
KEGG (Kanehisa et al., 2012; Kanehisa et al., 2018)	http://www.genome.jp/kegg	Kyoto Encyclopedia of Genes and Genomes
BioCyc (Karp et al., 2017)	http://biocyc.org	Collection of more than 17,000 Pathway/Genome Databases
BioPath (Reitz et al., 2004)	https://www.mn-am.com/databases/biopath	Biochemical molecules, reactions and pathways
Biochemical Pathways (Artimo et al., 2012)	http://biochemical-pathways.com	Digitized version of the Roche Applied Science 'Biochemical Pathways' wall chart
EAWAG-BBD (Gao et al., 2010)	http://umbbd.ethz.ch	EAWAG Biocatalysis/ Biodegradation Database
MetaNetEx.org (Moretti et al., 2016)	http://metanetx.org	Automated Model Construction and Genome Annotation for Large-Scale Metabolic Networks
Reactome (Fabregat et al., 2017; Jassal et al., 2019)	https://reactome.org	Manually curated and peer-reviewed pathway database

information (compound, glycan, reaction, enzyme); health information (network, variant, disease, drug, environ) and drug labels (medicus).

The *BioPath database* (REITZ et al., 2004) contains molecules, reactions, and biological pathways. Its first version is based on the Roche Applied Science 'Biochemical Pathways' wall chart and is extend with additional information from literature. In its version 3, BioPath contained about 14,000 chemical structures and about 3,900 biochemical transformations. A new version of the book form was also published (MICHAL et al., 2012).

The *EAWAG-BBD* contains information about microbial biocatalytic reactions and biodegradation pathways for xenobiotic compounds (GAO et al., 2010). Information on microbial enzyme-catalyzed reactions that are important for biotechnology can also be found. A Swiss bioinformatics group has opened their database for automated model construction and

genome annotation for large-scale metabolic networks, providing links to several hundred genome-scale metabolic networks (GANTER et al., 2013).

The *Reactome* database (FABREGAT et al., 2017; JASSAL et al., 2019) is a manually curated, peer-reviewed, open-source and open access pathway database. It contains information on signaling and metabolic molecules. The database features pathway visualization, data analysis tools as well as downloads of all data in major open-data standards such as Systems Biology Markup Language (SBML), Systems Biology Graphical Notation (SBGN) and Biological Pathway Exchange (BioPax) (DEMIR et al., 2010).

Organism-Specific Databases

Information about specific organisms are often collected in organism-specific databases (Table 2.6). Examples for such databases are *EcoCyc* (KESELER et al., 2017), which contains *E. coli* K-12 MG1655 data, or the *S. cerevisiae* Genome Database *SGD* (CHERRY et al., 2011).

Table 2.6: Organism specific databases

Database	URL	Content
EcoCyc (Keseler et al., 2017)	`http://ecocyc.org`	*E. coli* K-12 MG1655
SGD (Cherry et al., 2011)	`http://www.yeastgenome.org`	*S. cerevisiae* Genome Database

Summary

In this thesis, the data for the genome-scale metabolic network reconstructions and the organism-specific network reconstructions is exclusively taken from KEGG. However, the information contained in the databases listed in Tables 2.1 to 2.5 could additionally be used to extend and enhance the networks. For organism-specific networks, especially the information in the databases listed in Tables 2.4 and 2.6 are helpful. They already contain models of numerous organisms, which can either be used with minimal changes or taken as a basis for network reconstructions in combination with further resources.

2.2 Bioinformatic Tools

For the automated reconstruction of networks and analysis of network reconstructions, various bioinformatic tools are available. A selection is listed in Table 2.7.

Table 2.7: Bioinformatic tools for automated reconstruction of metabolic network models and their analysis

Database	URL	Content
Automated reconstruction		
Model SEED (Overbeek et al., 2005)	http://modelseed.org	A comparative genomics environment for curation of genomic data
Pathway Tools (Karp et al., 2015)	http://bioinformatics.ai.sri.com/ptools	A symbolic systems biology software system
KOBAS (Ai et al., 2018; Xie et al., 2011)	http://kobas.cbi.pku.edu.cn	KEGG Orthology Based Annotation System
GLAMM (Bates et al., 2011)	http://glamm.lbl.gov	Genome-Linked Application for Metabolic Maps
GEMSiRV (Liao et al., 2012)	http://sb.nhri.org.tw/GEMSiRV/en/GEMSiRV	Software platform for genome-scale metabolic models simulation, reconstruction, and visualization
ERGO	https://www.igenbio.com/ergo	Genome Analysis and Discovery System for the in silico analysis of organisms
Analysis		
CellNetAnalyzer (von Kamp et al., 2017)	http://www.mpi-magdeburg.mpg.de/projects/cna	MATLAB package for structural and functional analysis of biochemical networks
Metatool (von Kamp et al., 2006)	https://www.schleiden.uni-jena.de/Software.html	Computation of structural properties of biochemical reaction networks
Efmtool (Terzer et al., 2008)	http://www.csb.ethz.ch/tools/software/efmtool.html	Computation of elementary flux modes of metabolic networks
COBRA Toolbox (Heirendt et al., 2019)	https://opencobra.github.io	COnstraints-Based Reconstruction and Analysis

Model SEED (OVERBEEK et al., 2005) is a web-based resource for the high-throughput generation, optimization, and analysis of genome-scale metabolic network models. It integrates and augments technologies for the genome annotation, the construction of gene-protein reaction associations, the generation of biomass reactions, reaction network assembly, thermodynamic analysis of reaction reversibility, and model optimization to generate draft genome-scale metabolic network models. The generation of a metabolic network reconstruction from the assembled genome sequence takes about 48 h and automates nearly every step.

PathwayTools (KARP et al., 2015) is a software environment for the creation of pathway/genome databases (PGDBs) such as EcoCyc (KESELER et al., 2017). It allows the prediction of metabolic pathways and operons and network gap filling. Curators can interactively edit PGDBs. A large number of query and visualization tools as well as tools for comparative and systems biology analyses are available. Pathway Tools consists of three components. PathoLogic is used to create new PGDBs from annotated genomes. The Pathway/Genome Editors allow for the refinement of PGDBs. With the Pathway/Genome Navigator querying, visualization and analyses of PGDBs can be carried out.

Metatool (von KAMP et al., 2006) is a user-friendly tool for the calculation of elementary flux modes, conservation relations, and enzyme subsets in metabolic networks. Version 5.1 can be embedded into GNU Octave and MATLAB through script files and shared libraries. For calculations, the metabolic network data can be supplied to the program through the Metatool input format, as an SBML file or as stoichiometric matrix directly.

CellNetAnalyzer (von KAMP et al., 2017) is a MATLAB package with tools for the structural and functional analysis of different types of biochemical networks. For all computations, only the network topology is needed. CellNetAnalyzer allows the construction, input, and output of network projects via the Network Composer, text files, or SBML. Furthermore, it is possible to visualize network maps, either through import from KEGG or TRANSPATH or with external drawing tools. The functional network analysis covers the characterization of functional states of a network, the detection of functional dependencies, or qualitative predictions on effects of perturbations. For mass flow networks, there are two kinds of methods - namely constraint-based approaches and graph-theoretical analysis. Features are topological properties of the network such as dead-end metabolites, blocked or parallel reactions, and enzyme subsets. Metabolic flux analysis is also covered with the computation of steady-state flux distributions, feasibility check of flux scenarios, or optimal flux distributions for arbitrary linear objective functions. The computation of elementary modes for the metabolic path analysis is also possible. Minimal cut set analysis can help to detect strategies for the repression of certain

network functionality. From the graph-theoretical side, network properties such as shortest path lengths, connectivity of the network, or network diameter can be computed.

The Genome-linked Application for Metabolic Maps (GLAMM) (BATES et al., 2011) is a web interface unifying different tools for the reconstruction of metabolic networks from annotated genome data, visualization of metabolic networks together with experimental data, and investigation of the construction of novel transgenic pathways. GLAMM supports biological retrosynthesis and integration with tools of *MicrobesOnline*.

The genome analysis and discovery system *ERGO* developed by Integrated Genomics is a systems biology informatics toolkit for comparative genomics. With ERGO, one can capture, query, and visualize sequenced genomes and assign functions to genes, integrate genes into pathways, and identify unknown genes, gene products, and pathways. Its genomic database integrates with a collection of microbial metabolic and nonmetabolic pathways and proprietary algorithms. ERGO allows automated or manual annotation of genomes and genes, pathway analysis, multiple genome comparison, functional analysis of microarray data, data mining for the discovery of target genes, and *in silico* metabolic engineering and strain improvement.

2.3 Network Reconstruction

A metabolic network reconstruction is a structured database combining the available genetic, genomic, and biochemical data of an organism (REED et al., 2006). In general, a genome-scale metabolic network reconstruction consists of a list of reactions including their stoichiometry, the specific genes whose gene products are associated with these reactions, supporting annotation, and literature references. The fundamental goal of a network reconstruction is the accurate definition of the chemical transformations that take place among the chemical components of the network (REED et al., 2006). The construction and curation of a computational network links the organism's genome and expression to metabolic reaction fluxes, biomass, and energy production and consumption and enables the mathematical representation of the reactions and metabolic processes occurring in the organism. Metabolic networks can thus be used for *in silico* experiments (ZOMORRODI et al., 2012).

The process of compiling a (genome-scale) metabolic network can be broken up into five major stages (THIELE et al., 2010), as depicted in Figure 2.1. Briefly, in the first stage, a draft reconstruction of the network is built, which is refined in the second step. Then the

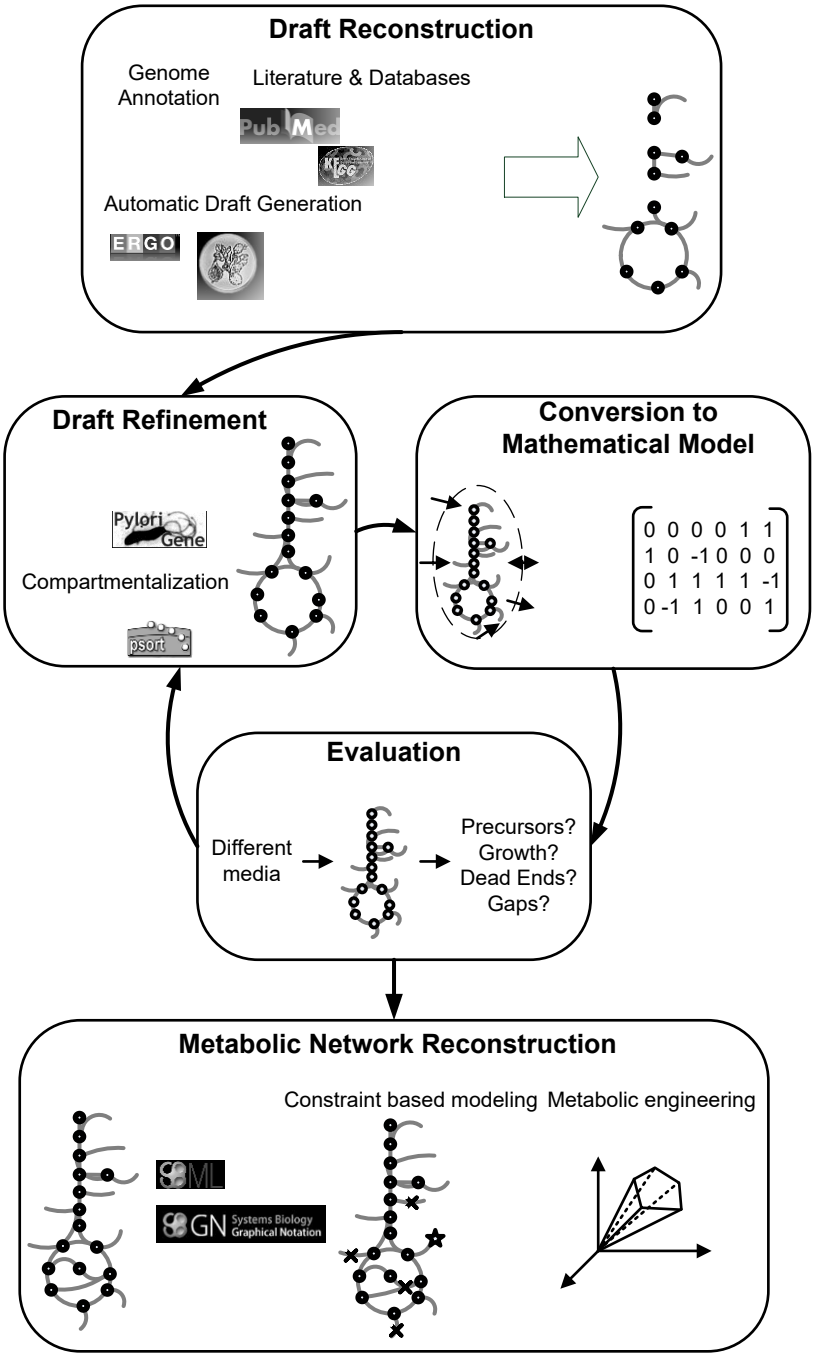


Figure 2.1: Workflow of an iterative network reconstruction process

network is converted to a mathematical model. In the fourth step, the reconstructed network is validated and can then be used for further experiments.

The first stage consists of creating a draft reconstruction of the network, at minimum containing a list of genes with their associated reactions and the corresponding EC numbers. The draft is based on the genome annotation for the most recent version of the target organism's genome and data from biological databases (Tables 2.1 to 2.5) and results in a collection of the genome-encoded functions of the metabolism. The important information for each gene is its function, its position in the genome and coding region, the strand and locus names, and the protein it codes for. For eukaryotes, information regarding alternative transcripts is also of interest because they may have distinct functions or a different cellular localization (THIELE et al., 2010). Candidate metabolic functions for the draft reconstructions can be retrieved by using GO categories, EC numbers, and biochemical databases (THIELE et al., 2010) (Tables 2.1 to 2.5). In general, the creation of a network draft is carried out automatically with the help of different software tools (Table 2.7). The automated genome annotation process with ERGO provides a draft annotation that requires manual curation to add organism specific information. The Basic Local Alignment Search Tool (BLAST) (ALTSCHUL et al., 1990) is used to annotate gene function based on orthology with other annotated genomes provided in online databases as well as phylogenetic approaches. Model SEED starts with an unannotated genome sequence and builds a draft metabolic network with gap filling and verification features (HAGGART et al., 2011). To refine this draft, a manual reconstruction refinement step is necessary.

The manual refinement step starts with an initial evaluation of the completeness of the draft reconstruction for identifying missing functions in the network. The draft can be reviewed pathway by pathway, starting from canonical pathways. The reactions of the model are evaluated in their metabolic context such that missing gene annotations and missing reactions can be identified easier. The use of network maps that show the environment of reactions is also convenient. Such maps can be found in databases such as KEGG or in organism-specific literature (THIELE et al., 2010). Correct stoichiometry requires complete balancing of elements and charges. Some databases may lack information on protons and water, for example. The incorporation of thermodynamic information is also of great value for the network model. Reaction directions can be based on the reaction's thermodynamic favorability, which can be determined from Gibbs free-energy changes. This information can be obtained from literature. However, the available data are rarely sufficient for genome-scale reconstructions, but rather for smaller models. Software, such as the biochemical thermodynamics calculator eQuilibrator (FLAMHOLZ et al., 2012), can be used to estimate thermodynamic parameters for

biological reactions in networks and pathways (NOOR et al., 2014) using the group contribution method (NOOR et al., 2013). If no thermodynamic information is available, the reaction should be left reversible. Organism-specific functions should also be taken into account, such as the use of substrate and cofactors, which can differ between different organisms. The review of primary literature dealing with the metabolism and function of the organism is necessary to identify these organism-specific characteristics. For a growing number of organisms, specific textbooks exist, which are a resource for additional information. When organism-specific information is not available to the desired extent, data from phylogenetic neighbors can be taken into account. Gene-protein reaction relationships, which connect the genes with their associated enzymes via Boolean logic, allow the simulation of phenotypic effects of gene knockouts. Also, the compartmentalization information for metabolite and reaction localization as well as intracellular transport reactions have to be checked. If no sufficient data are available, the respective proteins should be assumed to reside in the cytosol. However, incorrect assignment can lead to additional network gaps (THIELE et al., 2010). For cell-free systems, e.g. cell lysates, there exist no compartments. A network reconstruction for such a system thus has no need for compartmentalization, intracellular transport reactions and separated metabolite pools. They can be neglected in the reconstruction process.

Furthermore, biomass composition, maintenance parameters, and growth conditions of the organism are to be determined by different experimental and computational methods. When designing a biosynthetic synthesis pathway, the selection of substrates and cofactors has to take into account toxicity for the host organism. For living organisms, non-toxic substrates and cofactors should be preferred. This does not need to be incorporated into network reconstructions for cell-free systems, where reactants can also be present in non-physiological concentrations.

The conversion from the network reconstruction to a mathematical model for validation and *in silico* applications consists of three steps, which can mostly be automated with suitable tools. The first step is the mathematical representation of the network as a stoichiometric matrix. In the second step, the boundaries of the system are defined. For each metabolite that can be consumed or secreted, an exchange reaction is added to allow the definition of environmental conditions for *in silico* simulations. This is only needed for network reconstructions of living organisms and is omitted in network reconstructions of cell-free systems. For network reconstructions of living organisms, constraints are added to the model to turn it into a condition-specific model. Thermodynamic data for enzyme capacities or regulation help to determine a set of feasible steady-state flux solutions.

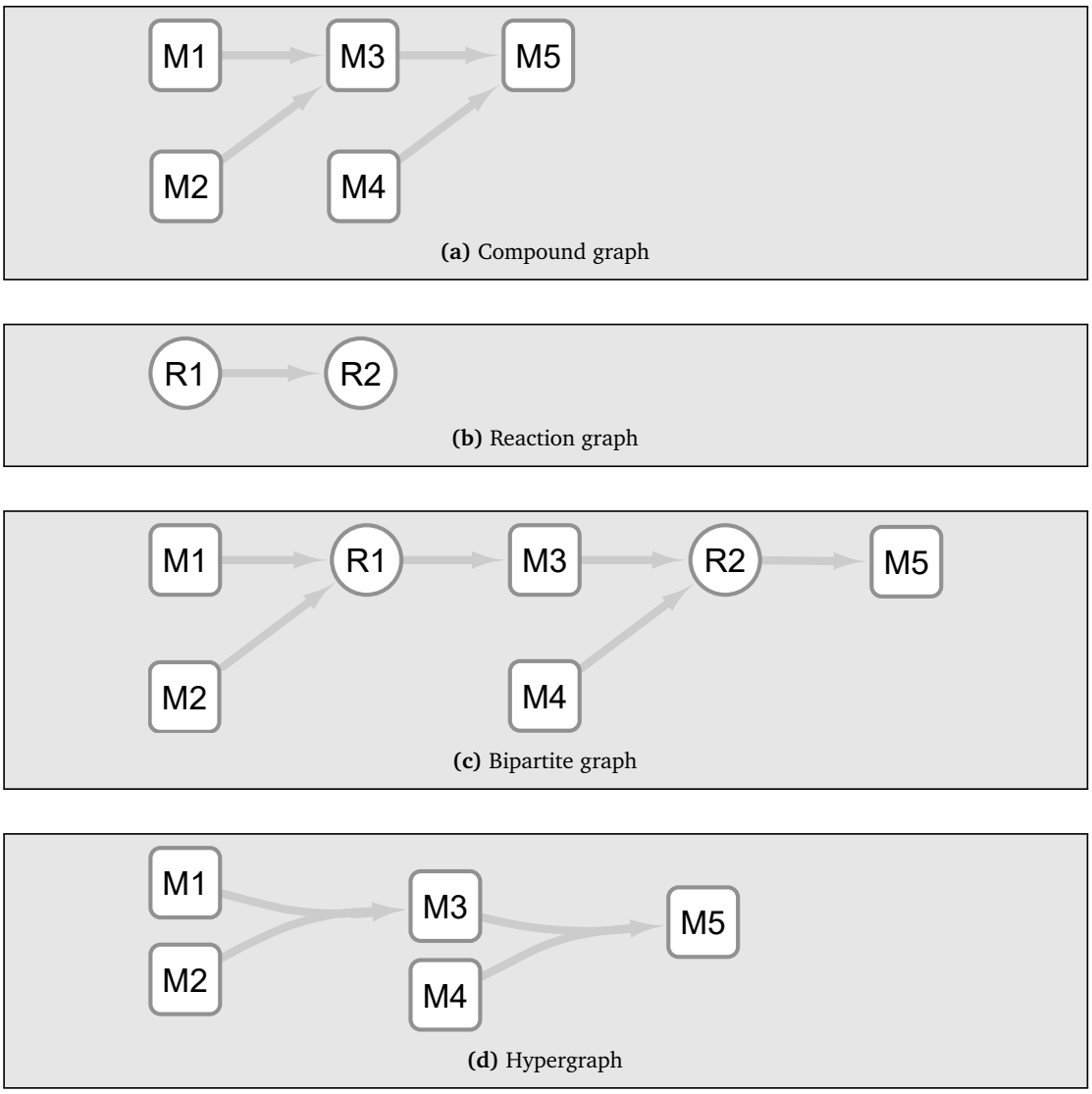
The evaluation stage includes network verification, evaluation, and validation steps to help detect gaps in the network. To find candidates for filling gaps, an intensive literature search is needed that helps to identify the environment of the dead-end metabolites. Databases such as ATLAS of Biochemistry (HADADI et al., 2016) and other cheminformatic tools (e.g. the rePrime procedure (KUMAR et al., 2018)) can also help to close network gaps with *de novo* reactions. Methods like BridgIT (HADADI et al., 2019) can be used to identify genes and proteins for orphan reactions that are not associated with enzymes.

For network reconstructions of living organisms one must also take care of stoichiometrically balanced cycles formed by internal network reactions that can carry fluxes despite closed exchange reactions (THIELE et al., 2010). The model must be tested for its ability to synthesize all biomass precursors, such as amino acids, nucleotide triphosphates, or lipids with different medium compositions. This can be done by growing the organism on specific carbon sources (HAGGART et al., 2011). It should be checked if the model could reproduce known incapabilities of the organism. It is also advised to compare the predicted physiological properties with known properties such as carbon splits in the central metabolic pathways of the organism. For cell-free network reconstructions, these steps can be omitted.

Network reconstructions can be used for several major applications that address different aims of these models (OBERHARDT et al., 2009), such as using metabolic network reconstructions for putting high-throughput experimental data into context. They can also be used for discovery of network properties, hypothesis-driven discovery, and exploration of multispecies relationships. Network reconstructions also have applications in metabolic engineering, where they can be used for constraint-based modeling and the *in silico* prediction of possible cellular phenotypes without the need for kinetic data. The main concept behind network metabolic modeling is the identification and mathematical definition of constraints for the separation of feasible and infeasible metabolic behavior. These constraints are usually much easier to identify than kinetic parameters. There are three types of constraints: Physicochemical constraints deal with mass and energy conservation, the dependency of reaction rates on metabolite concentrations, and the negative free-energy change for spontaneous reactions. Environmental constraints are imposed as a result of specific conditions such as nutrients, whereas regulatory constraints express the effects of gene expression and enzyme activity regulation properties.

2.4 Network Representation

Metabolic network models can formally be described as graphs G(V, E), where V is the set of vertices and E is the set of edges connecting node pairs. In a directed graph, the edges are ordered, whereas in an undirected graph an edge is represented by an unordered node pair. There are different possibilities to represent a metabolic network graph. In a *compound*



(a) Compound graph

(b) Reaction graph

(c) Bipartite graph

(d) Hypergraph

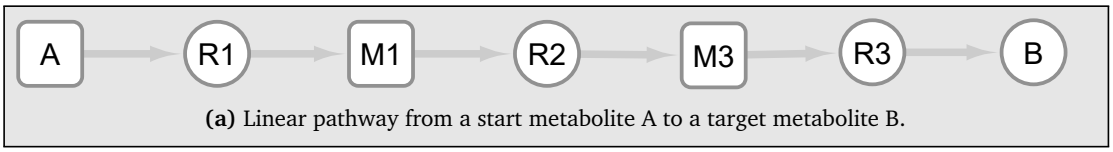
Figure 2.2: Graph representations. Rounded squares: metabolites; circles: reactions.

graph (Figure 2.2(a)), the nodes represent the chemical compounds. An (un)directed edge connects two compounds if they are substrate and product of the same reaction. The dual

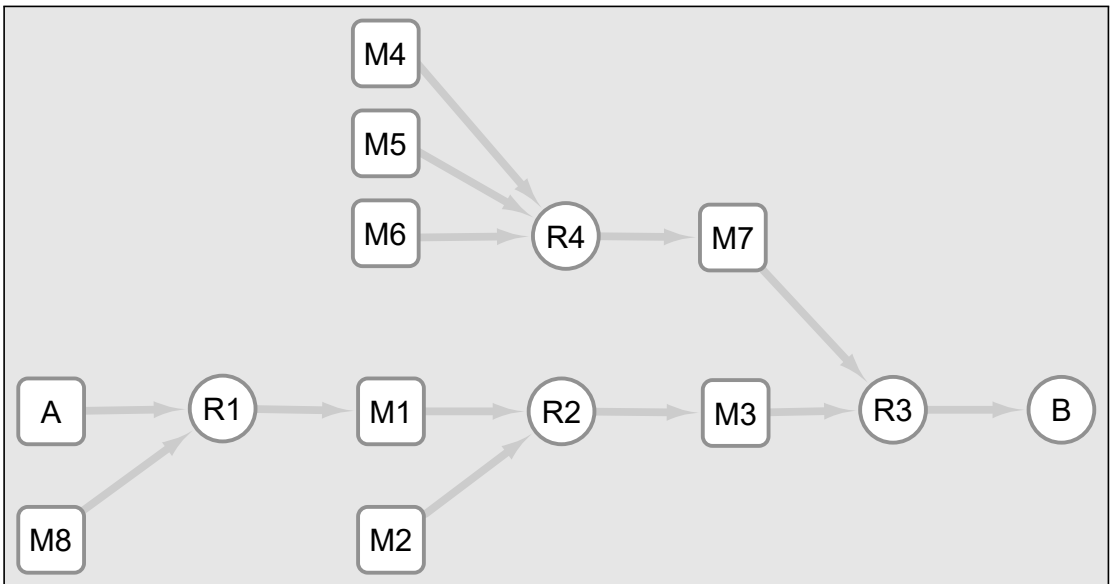
form of the compound graph is the *reaction graph* (Figure 2.2(b)), where the nodes represent reactions. Edges in this graph connect two reactions if one reaction has products that are substrates of the other reaction. Both types of representation have similar limitations, as they are both ambiguous and do not represent all information of the network. Another type of graph representation is the *bipartite graph* (Figure 2.2(c)), in which there are two classes of nodes representing compounds and reactions in the graph. Directed or undirected edges in a bipartite graph are only possible between two nodes of different classes. A substrate is defined by a directed edge from a compound to a reaction node, a product by an edge from a reaction to a compound node. An equivalent representation of a bipartite graph and also the generalization of a compound graph is a *hypergraph* (Figure 2.2(d)) with directed or undirected edges. In such a graph, a hyperedge relates a set of substrates to a set of products. This graph type allows an unambiguous representation of reactions and compounds, but it has limited coverage because reaction control factors cannot be represented. Diverse graph types and data models for biochemical pathways are reviewed in more detail in (DEVILLE et al., 2003).

A biological pathway can be defined in several different ways, depending on the underlying biological network representation or the context in which the pathway is considered. When metabolism is defined as a network of chemical reactions catalyzed by enzymes and connected by substrates and products, then the most basic definition of a metabolic pathway is a coordinated series of reactions (DEVILLE et al., 2003). Given a compound graph, a pathway can be a sequence of metabolites that are linked by reactions or substrate-product pairs. In the simplest case, the sequence is a linear path from a start metabolite A to a target metabolite B (Figure 2.3(a)). In most cases, this definition is too basic. Especially in the context of biological synthesis, this kind of pathway does not cover all information needed. It is much more meaningful to look at branched pathways. A branched pathway to a given target metabolite does not have a single start metabolite, but it rather can have multiple start metabolites by taking into account every substrate of each reaction involved in the pathway to produce the target metabolite B (Figure 2.3(b)).

A basic mathematical representation of metabolic networks is the stoichiometric matrix. It represents its charge and elementary balanced metabolic reactions and thus quantifies the stoichiometric relationship between the metabolites in a reaction. The rows and columns of the matrix correspond to the metabolites and reactions of the network. Its nonzero elements are the stoichiometric coefficients, which are positive for products and negative for substrates. For genome-scale metabolic networks, the stoichiometric matrix is sparse because relatively few metabolites participate in a given reaction.



(a) Linear pathway from a start metabolite A to a target metabolite B.



(b) Branched pathway to a target metabolite B.

Figure 2.3: Pathway definitions. Rounded squares: metabolites; circles: reactions.

Another quite simple representation method for metabolic networks is compiling all information in spreadsheets. The spreadsheet should contain all gene names and their abbreviations. For each reaction, the reactant, substrate, product symbols, balanced stoichiometry, reversibility, compartment, associated protein, and its EC number should be included. Also of importance are literature references and a confidence rating for each annotation entry as well as comments.

SBML is a machine-readable format for representing biological models. Its basic idea is to cast a network reconstruction into a formal, computable form, thus allowing network analysis using simulations and other mathematical methods.

The SBGN (LE NOVERE et al., 2009) is a project that aims to standardize the graphical notation used in maps of biological processes. Currently, there are three different languages for different types of network maps. The Process Description (PD) language (ROUGNY et al., 2019) can be used to depict temporal courses of biochemical interactions of a network.

Relationships between entities of a network can be modeled with the Entity Relationship (ER) language (SOROKIN et al., 2015). The Activity Flow (AF) language (MI et al., 2015) visualizes the information flow between biochemical entities in the network, such as for representation of the effects of perturbations on the network.

BioPax (DEMIR et al., 2010) is an open standard language for the integration, exchange and visualization of biological pathways.

2.5 Metabolic Network Design and Manipulation

The planning of a biochemical synthesis in permeabilized cells involves primarily finding a synthesis route starting from available, inexpensive, and stable substrates. The overall goal is to obtain high product concentration with high yield and selectivity in the shortest possible time. High yield and selectivity can only be obtained if undesired side reactions do not take place. The products of undesired side reactions will also complicate downstream processing. Side reactions do not occur if one of the substrates required is missing. The substrate composition is a design variable when using permeabilized cells as biocatalysts. Side reactions can be eliminated by the deletion of the corresponding gene or by the addition of a selective inhibitor. A major engineering task of biosynthesis in permeabilized cells is the regeneration of cofactors, such as NAD(P)H (LEE et al., 2013) or ATP (HORINOUCHI et al., 2006; HORINOUCHI et al., 2012).

Metabolic engineering is the manipulation of enzymatic, transport, and regulatory functions of a cell through recombinant DNA technologies. One of its important objectives is the improvement of the cellular phenotype or the yield of a desired product. Traditionally, this is done by rationally selected gene deletions or overexpression of native and heterologous genes in an organism. To remove undesirable metabolic pathways in an organism, site-directed mutagenesis or homologous recombination can be used. To increase biochemical yields and add new functions, heterologous genes or even complete pathways can be introduced into the organism. *In silico* metabolic models allow rational predictions of the phenotypical response of changes in culture media, gene knockouts, and the incorporation of heterologous enzymes and pathways into an organism (BLAZECK et al., 2010).

Flux balance analysis is a widely used constraint-based method in metabolic engineering for studying biochemical networks. It allows for the *in silico* prediction of flux profiles that optimize a cellular objective, depending on the problem. Often, the biomass production or the production rate of a certain metabolite of interest is used as an objective. The fundamental

assumption for flux balance analysis is that the metabolism in the cell is at steady state as well as all reaction fluxes and metabolite concentrations (HAGGART et al., 2011). The input for flux balance analysis is the mathematical representation of the metabolic network as a stoichiometric matrix. The stoichiometric coefficients in the matrix constrain the flow of metabolites through the network. These steady-state mass balance equations for each metabolite and the environmental and growth conditions can be described mathematically in the form of constraints for an optimization problem. The metabolite balance equation is a homogeneous system of linear equations $S \cdot v(t) = 0$, where S is the stoichiometric matrix and $v(t)$ is the vector of reaction rates. It requires that each metabolite is consumed at the same rate as it is produced (TERZER et al., 2009). For the quantitative determination of the metabolic fluxes, linear programming can be used to solve the stoichiometric matrix for a given objective function under various constraints (T. Y. KIM et al., 2012). A linear program is a mathematical optimization model which requires maximizing or minimizing a given objective under a finite set of given constraints. The constraints describe the space of all eligible possibilities from which an optimal solution can be selected. They are generally given in the form of equalities and/or inequalities. In its canonical form, a linear program it can be formulated as

maximize	$\mathbf{c}^{\mathrm{T}}\mathbf{x}$
subject to	$A\mathbf{x} \leq \mathbf{b}$
and	$\mathbf{x} \geq \mathbf{0}$

where $\mathbf{x}$ is the vector of variables; $\mathbf{c}$ and $\mathbf{b}$ are vectors of known coefficients; and $\mathbf{A}$ is a known matrix of coefficients. $(\cdot)^{\mathrm{T}}$ is the matrix transpose. A feasible solution of the linear program is any assignment of $\mathbf{x}$ satisfying all constraints (CORMEN et al., 2009). For an integer linear program all variables are required to be integers, while in a MILP also non-integer variables are allowed. Finding a feasible solution for a linear problem can be generally done in polyomial time, for example with the simplex algorithm (CORMEN et al., 2009). However, finding a feasible solution of a integer linear problem is NP-hard, so there are no polynomial-time algorithms known. Nevertheless, commercially available tools such as the IBM CPLEX Optimizer as well as non-commercial solvers allow for the fast computation of a feasible solution. Note that in the worst case the running time is not polynomial.

Constraint-based methods focus only on reaction fluxes, neglecting enzyme kinetics and regulations that can influence the actual fluxes. Therefore, they have some limitations in their predictive capabilities (DUROT et al., 2009). However, they can be computed very

efficiently, even for large networks. Elementary mode analysis is an important method for metabolic network studies (von KAMP et al., 2006), allowing the enumeration of all independent minimal pathways in the network that are stoichiometrically and thermodynamically feasible. Elementary flux modes are independent flux distributions of a metabolic network at steady state. The inputs for elementary mode analysis are the reaction stoichiometries and reversibilities. All metabolites in the network are classified as either internal or external. The internal metabolites are balanced and the external metabolites are assumed to be buffered. The computation of elementary modes in large networks is difficult due to its combinatorial complexity (KLAMT et al., 2007). Once elementary modes are computed, the deletions necessary for the elimination of undesired side product formation can be directly identified. It requires that each metabolite is consumed at the same rate as it is produced (REED et al., 2006). For the quantitative determination of the metabolic fluxes linear programming can be used to solve the stoichiometric matrix for a given objective function under various constraints (ZOMORRODI et al., 2012). The constraints of the problem describe the space of all eligible possibilities from which an optimal solution can be selected. It has been shown that it is even possible to directly identify successful targets for the overexpression of enzymes just based on the known stoichiometry of a network (BOGMAN et al., 2003; HAGGART et al., 2011; NEUNER et al., 2011; RYAN et al., 1991).

However, these methods are not directly applicable for permeabilized cells because they assume a steady-state in the network. This will be discussed in detail in Section 4.1 Background where a different approach tailored for cell-free systems is presented.

CHAPTER 3

Aims and Scope

This thesis has two main aims. The first aim is to design a comprehensive method for finding and analyzing pathway candidates for synthesis in genome-scale metabolic network reconstructions of cell-free systems. The second aim of this thesis is to develop a method for creating and characterizing such network reconstructions from KEGG data for the planning of biosynthetic production pathways using cell-free systems.

In Part II Path-Finding and Network Analysis for Multi-Enzyme Biocatalysis, a newly developed method for finding pathways in a genome-scale metabolic network reconstruction is discussed. The method is based on a MILP and combines both topology of the graph based on the network and the stoichiometry of the reactions in the network model. The algorithm only requires the specification of the target product of interest to find pathways starting from arbitrary substrates and a set of ubiquitous cofactors (e.g. nucleoside triphosphates such as ATP; or NAD(P)H) and inorganic metabolites such as water or CO_2. A set of different ranking criteria to help finding well-designed and meaningful synthesis pathway candidates is also developed. These criteria include pathway length, reaction thermodynamics, the number of heterologous enzymes for a given host organism or cofactor requirement of the pathway, amongst others. The features of the method are presented using geranyl pyrophosphate (GPP), amygdalin, pyrrolysine and (S)-2-phenyloxirane as examples for target metabolites.

In Part III Network Reconstructions for Cell-Free Systems a method for the reconstruction of genome-scale metabolic network models based on KEGG data is presented and discussed. Not suitable data, e.g. generic metabolites, general reactions and reactions with invalid stoichiometry are removed with a filtering scheme. The network models also comprise stoichiometric and thermodynamic data that allow the definition of constraints to identify

potential pathways. A pan-organism network reconstruction containing all suitable reactions in KEGG is assembled. Furthermore, single organism network reconstructions from several organisms important in biotechnological production and scientific research, such as *Escherichia coli* and *Corynebacterium glutamicum*, amongst others, are established. These network models are analyzed with the help of graph theoretical methods to identify a set of metabolites that are potentially reachable from a defined set of starting metabolites. It is discussed how they can be used for the planning of biosynthetic production pathways. The usage of the path-finding tool is presented using the example of UDP-glucose as a target.

The workflow for network reconstruction and analyzing together with the path-finding method describes a powerful and highly customizable toolbox usable for the design of multi-enzyme biosynthetic production pathways. The data resulting from the studies presented in this work can be directly applied to the planning of biosynthetic production pathways and can also help setting up custom network reconstructions or improving existing network models.

Part II

Path-Finding and Network Analysis for Multi-Enzyme Biocatalysis

CHAPTER 4

Network Design and Analysis for Multi-Enzyme Biocatalysis

The following chapter is based on the published research article

> BLASS, L. K., WEYLER, C., & HEINZLE, E. (2017): Network design and analysis for multi-enzyme biocatalysis. *BMC Bioinformatics* (Aug. 2017), vol. *18*([1]): 366. https://doi.org/10.1186/s12859-017-1773-y

that I co-authored.

Appendix A is based on the the supplementary material of this research article.

Abstract

As more and more biological reaction data becomes available, the full exploration of the enzymatic potential for the synthesis of valuable products opens up exciting new opportunities but is becoming increasingly complex. The manual design of multi-step biosynthesis routes involving enzymes from different organisms is very challenging. To harness the full enzymatic potential, we developed a computational tool for the directed design of biosynthetic production pathways for multi-step catalysis with *in vitro* enzyme cascades, cell hydrolysates and permeabilized cells.

We present a method which encompasses the reconstruction of a genome-scale pan-organism metabolic network, path-finding and the ranking of the resulting pathway candidates for proposing suitable synthesis pathways. The network is based on reaction and reaction pair data from the KEGG and the thermodynamics calculator eQuilibrator. The pan-organism

network is especially useful for finding the most suitable pathway to a target metabolite from a thermodynamic or economic standpoint. However, our method can be used with any network reconstruction, e.g. for a specific organism. We implemented a path-finding algorithm based on a MILP which takes into account both topology and stoichiometry of the underlying network. Unlike other methods we do not specify a single starting metabolite, but our algorithm searches for pathways starting from arbitrary start metabolites to a target product of interest. Using a set of biochemical ranking criteria including pathway length, thermodynamics and other biological characteristics such as number of heterologous enzymes or cofactor requirement, it is possible to obtain well-designed meaningful pathway alternatives. In addition, a thermodynamic profile, the overall reactant balance and potential side reactions as well as an SBML file for visualization are generated for each pathway alternative.

We present an *in silico* tool for the design of multi-enzyme biosynthetic production pathways starting from a pan-organism network. The method is highly customizable and each module can be adapted to the focus of the project at hand. This method is directly applicable for (i) *in vitro* enzyme cascades, (ii) cell hydrolysates and (iii) permeabilized cells.

4.1 Background

While thousands of enzymes are already known, numerous new enzymes or new enzymatic activities are still discovered every year. Many of these biocatalysts accept multiple substrates and even catalyze different reactions. From a biotechnological point of view, the enzymatic potential of nature can be considered an extremely versatile tool potentially giving access to countless valuable products ranging from bulk chemicals to most complex drug compounds. The methods for such syntheses can range from using single isolated enzymes over multi-enzyme systems or enzyme cascades up to syntheses with cell lysates or permeabilized cells (HEINZLE et al., 2013).

However, the full exploration of the enzymatic potential is often hampered by the sheer amount and complexity of available reaction data. When manually designing a multi-step synthesis route to a certain metabolic intermediate, the network of alternative synthesis pathways quickly grows highly complex as more reaction steps are introduced. Additionally, assembling all reactions that lead to each reactant is extremely time consuming. The manual determination of the most suitable pathway candidate is challenging as multiple aspects such as thermodynamics, cofactor use, etc. need to be considered. To more easily harness the full potential of the enzymatic toolbox we developed a computational tool for the directed

design of biosynthetic production pathways for interesting products in cell extracts and permeabilized cells.

The search for pathways in genome-scale metabolic networks is a common task of wide interest and there is a large variety of path-finding and pathway design methods. Most of those methods can be categorized into one of two types, namely stoichiometric methods and graph-based methods. Stoichiometric methods make use of the stoichiometry of a network to analyze the metabolism under the assumption of a steady-state condition. Popular and mathematically well understood methods are for example elementary flux modes (SCHUSTER et al., 1999) or flux balance analysis (PHARKYA et al., 2004; TERVO et al., 2016). Graph-based methods in general neglect stoichiometry and treat the networks as graphs in a mathematical sense and search for pathways based on connectivity (FAUST et al., 2009), with the use of atom or atom group tracking (BLUM et al., 2008a, 2008b; HUANG et al., 2017), retrosynthesis (CARBONELL et al., 2011; HATZIMANIKATIS et al., 2005), heuristic search algorithms (MCCLYMONT et al., 2013) or evolutionary algorithms (GERARD et al., 2015). In the last years, methods combining stoichiometry and structural properties of networks emerged, e.g. the so called carbon flux paths proposed by Pey et al. (PEY et al., 2014; PEY et al., 2011). However, the majority of these methods tackles the problem of finding pathways between two given metabolites and does not take into account a search starting with an arbitrary metabolite in the network. Another drawback of these methods for our focus of application is that most of them assume a steady-state condition for the major part of the network. This is valid for living cells or cells with intact membranes. In these cases the actual reactions are running in a cellular compartment that keeps all intermediates separated from the bioreactor, whereas in the case of enzyme cocktails and permeabilized cells the reaction compartment is identical to the bioreactor used. Examples of the latter type of reaction systems are becoming increasingly popular (CARSTEN et al., 2015; DUDLEY et al., 2016; KARIM et al., 2016; KOIZUMI, 2003; KOIZUMI et al., 2000; KOIZUMI et al., 1998; KRAUSER et al., 2015; WEYLER et al., 2015).

We thus propose a tool which encompasses the reconstruction of a genome-scale pan-organism metabolic network, the implementation of a path-finding algorithm and the ranking of pathway candidates for proposing suitable synthesis pathways starting from arbitrary substrates.

4.2 Methods

In the following we will present the individual parts of our method. Figure 4.1 shows the workflow through its different components.

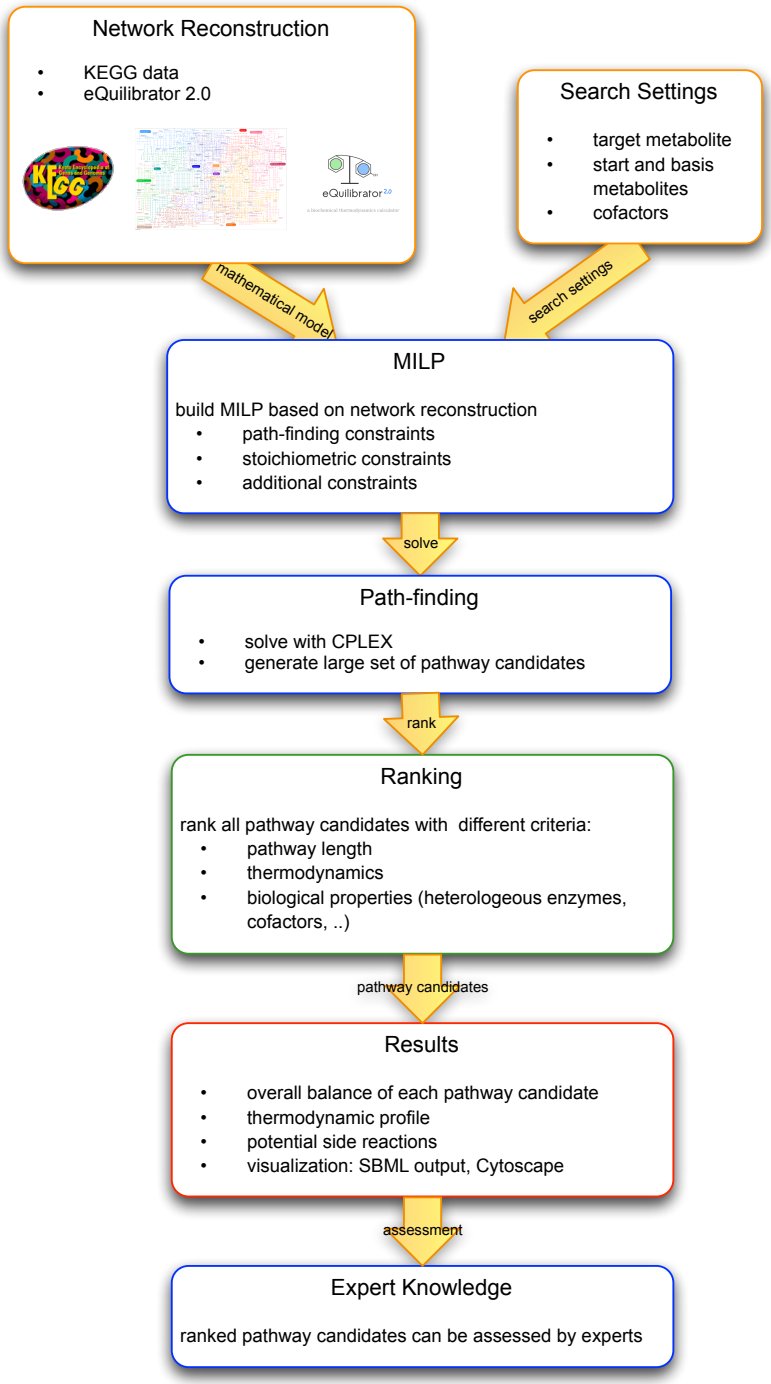


Figure 4.1: Workflow through the components of our tool. We start with a network reconstruction which is then used for path-finding with the presented MILP. The resulting pathway candidates are ranked according to the different ranking criteria.

The first step is the network reconstruction where the network is built with data from KEGG (KANEHISA et al., 2000; KANEHISA, SATO, KAWASHIMA, et al., 2016) and the biochemical thermodynamics calculator eQuilibrator 2.0 (FLAMHOLZ et al., 2012; NOOR et al., 2013). Details on how the network is compiled are given in Section 4.2 Network Reconstruction. The path-finding in the network is based on an optimization algorithm developed by Pey et al. (PEY et al., 2011). It combines graph-based path-finding and reaction stoichiometry in a MILP. The algorithm with our extensions is presented in detail in Section 4.2 Mathematical Model. In a further stage the resulting pathway candidates are ranked using different criteria. We will give details on the ranking in Section 4.2 Filtering and Ranking. The output is a list of ranked pathway candidates which can be assessed with expert knowledge to help determining the most suitable synthesis pathway for a desired product.

Network Reconstruction

We combine data from different KEGG databases and eQuilibrator 2.0 for the reconstruction of a pan-organism network with data from all organisms contained in KEGG release 78.1 from May 1, 2016.

Reaction and Reaction Pair Data

The reaction network was reconstructed with COnstraint-Based Reconstruction and Analysis (COBRA) Toolbox (SCHELLENBERGER et al., 2011) using reactions from KEGG REACTION. We excluded reactions with the comments 'generic' and 'incomplete' in their data entries; reactions with ambiguous stoichiometry with stoichiometric coefficient n in the reaction equation; as well as reactions involving glycans with G numbers in KEGG.

From all remaining reactions in the model we built a network of reaction pairs, the so called arcs. A reaction pair is a biologically meaningful substrate-product pair in a reaction. We derived the arcs from the KEGG RPAIR database [1] containing reaction pairs for each reaction. The reaction pairs in KEGG are classified into five categories (KOTERA, HATTORI, et al., 2004) from which we used the main-pairs, describing the main changes on the substrates in a reaction and the trans-pairs which describe transferase reactions. We did not use the remaining three types cofac-pairs, ligase-pairs and leave-pairs. However, they can be included at user's discretion.

1 discontinued since KEGG release 80.0, October 1, 2016

Our network reconstruction comprises a total of 9038 reactions (10160 including reversible reactions), 7405 metabolites and 14803 arcs.

Thermodynamic Data

The KEGG REACTION database does not contain any detailed information about reaction directions, so we incorporated thermodynamic data from the biochemical thermodynamics calculator eQuilibrator 2.0. The component contribution method used (NOOR et al., 2013) provides different types of the reaction Gibbs energy. $\Delta_r G'^{\circ}$ expresses the change of the Gibbs free energy of a reaction at a given pH and ionic strength I in 1 M concentration of the reactants. However, for metabolic reactions in cells it makes more sense to use physiologically meaningful concentrations. For $\Delta_r G'^{m}$ the concentration of the reactants is thus set to 1 mM. For all calculations standard parameters are used which are a temperature of 25 °C (298.15 Kelvin), a pH of 7 and a pressure of 1 bar. We set the threshold for the discrimination of reversible and irreversible to $\Delta_r G = 15$ kJ/mol. Reactions without available thermodynamic data are considered irreversible in the direction given in the reaction equation from KEGG.

Network Details

We categorize the metabolites in the model into different sets which we treat differently in our path-finding method. All sets are given in Additional file 1 of (BLASS et al., 2017). A Venn diagram of these sets is depicted in Figure 4.2.

As *start metabolites* S we denote all metabolites that can be potential start points of a metabolite path. A metabolite path is a sequence of metabolites through the network connected by arcs. We compiled the list of possible start metabolites with all metabolites in the model contained in arcs with a molecular mass between 0 and 300. A subset of the start metabolites are the so called *basis metabolites* B. They are an expert-curated set of metabolites that are hubs of the arc network, easily available and inexpensive, such as D-glucose (C00031[1]) or pyruvate (C00022).

As *cofactors* we denote metabolites that are required for the activity of the enzymes catalyzing the reactions in the network but are not directly part of the reaction chain. We exclude arcs containing cofactors from the set of arcs to prevent biologically meaningless shortcuts in the network. The list is expert-curated and contains mono-, di- and triphosphates (e.g.

1 KEGG compound ID

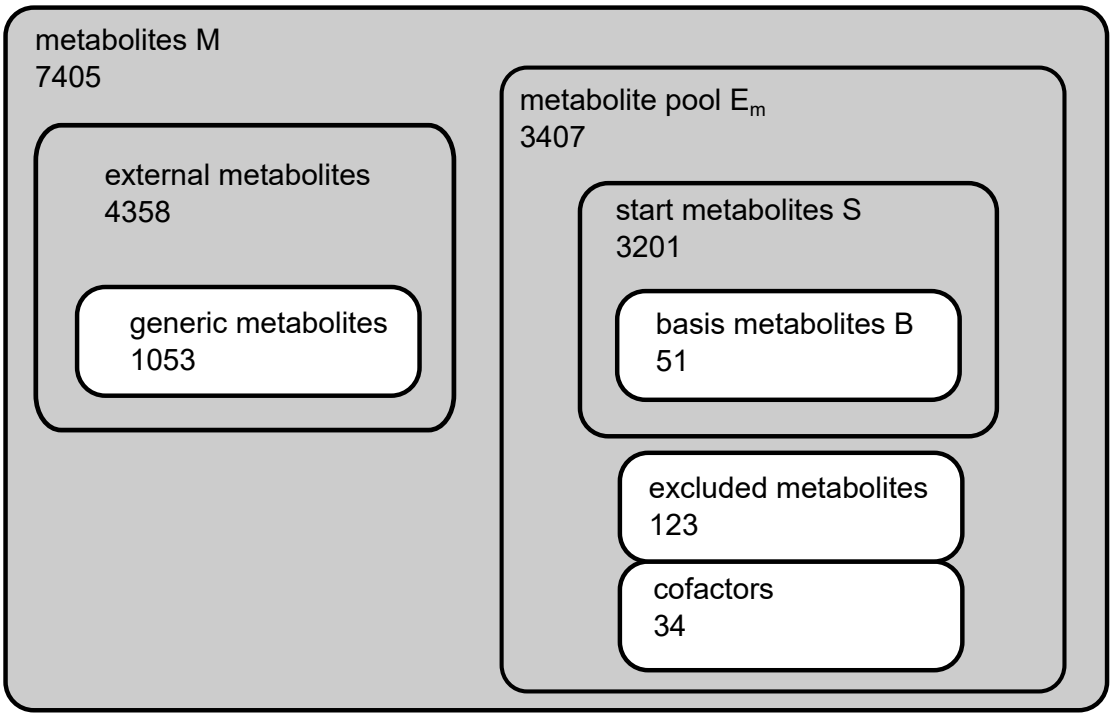


Figure 4.2: Venn diagram with the different metabolite categories in the network reconstruction. Metabolites M: all metabolites in the network; metabolite pool E_m: metabolites considered available from start; start metabolites: all metabolites in the model contained in arcs with a molecular mass between 0 and 300; basis metabolites: expert-curated subset of start metabolites; cofactors: cofactors for enzymes; excluded metabolites: treated as cofactors; external metabolites: not contained in the metabolite pool, cannot be externally supplied; generic metabolites: marked as 'generic' in their KEGG entry; the grey background indicates the set that can contain the product P.

adenosine 5'-monophosphate (AMP) (C00020), adenosine 5'-diphosphate (ADP) (C00008) and ATP (C00002)), electron carriers such as nicotinamide adenine dinucleotide (NAD^+) (C00003) and others. The mono- and diphosphates are usually not considered cofactors, but we chose to incorporate them into the list to avoid unnecessary interconversions between them on the pathway candidates. The set of *excluded metabolites* is treated in the same way as the cofactors. It contains metabolites that are considered as freely available, such as water, oxygen or CO_2. As the *metabolite pool* E_m we denote the superset of metabolites we consider as freely available. This set consists of start metabolites, basis metabolites, cofactors and excluded metabolites. As *external metabolites* we denote all metabolites that are not contained in the metabolite pool. They have to be produced in a production pathway and cannot be externally supplied. *Generic metabolites* are metabolites that are marked as 'generic'

in their KEGG entry, such as peptide (C00012) or protein (C00017). In our network we treat them as external metabolites and exclude arcs containing those metabolites from the arc network. The pool of external metabolites also contains metabolites with arcs that are not start metabolites as well as all other metabolites that are not part of any other set.

Path-Finding

In the following we introduce our method for finding pathway candidates in the network by means of a MILP.

Mathematical Model

Given a metabolic model with the set of reactions R and the set of metabolites M we build the network of arcs. We also use the |M|-by-|R| stoichiometric matrix of the network, where each row corresponds to a metabolite and each column corresponds to a reaction. An entry in the matrix represents the stoichiometric value of a metabolite in the respective reaction, where negative values indicate a reactant and positive values indicate a product. Reversible reactions appear in the model as two different reactions with opposite directions.

MILP

The algorithm presented is based on an algorithm proposed by Pey et al. (PEY et al., 2011). However, in comparison to the original algorithm we changed the problem statement. Pey et al. dealt with the question of finding the K-shortest flux paths between a given source and a target metabolite. Different from this problem statement we do not specify any specific starting metabolite, but our algorithm identifies suitable starting metabolites for finding a pathway to a target metabolite P. In our definition, a *pathway* consists of two parts. The first part is a sequence of metabolites connected by reactions. It starts with a reaction that has one of the possible start metabolites as substrate and ends with a reaction with the desired target metabolite as a product. This part is called the *linear path*. The second part is a minimal set of reactions supplying substrates that are needed by the reactions on the path which are not contained in the metabolite pool. These are called *supplying reactions*.

We introduce the set of binary variables u_{ij} which are 1, if an arc from i to j is part of the linear path, and 0 otherwise (for $i,j = 1,\ldots,|M|$). The first constraint given by equation (4.1) establishes that there is exactly one arc on the linear path ending in the target metabolite P,

whereas the second constraint in equation (4.2) assures that no arc on the linear path starts with P.

$$\sum_{i=1}^{|M|} u_{iP} = 1 \tag{4.1}$$

$$\sum_{j=1}^{|M|} u_{Pj} = 0 \tag{4.2}$$

The two constraints ensure that the target P is always the last node on each identified path and thus the path actually ends with the desired product. Both constraints have been adopted from (PEY et al., 2011). Inequality (4.3) states that the number of arcs entering a node l from the set of possible start nodes S on the path is smaller or equal to the number of arcs leaving it.

$$\sum_{i=1}^{|M|} u_{il} \leq \sum_{j=1}^{|M|} u_{lj} \quad l \in S; \quad l \neq P \tag{4.3}$$

This means that a metabolite l is either the starting metabolite of a path ($\sum u_{il} = 0$ and $\sum u_{lj} = 1$) or the metabolite is an intermediate ($\sum u_{il} = \sum u_{lj}$). In the trivial case where l is not on the path, both sums are zero. The idea of the constraint has been adopted from (PEY et al., 2011). However, we changed it to incorporate the set of starting metabolites, which has not been introduced in the original MILP. For the set of basis metabolites B we introduce a constraint formulated in equation (4.4) stating that the number of arcs entering a node l from the set of basis metabolites B should be zero. This means that a basis metabolite can only appear as the first metabolite in a metabolite path and not as an intermediate.

$$\sum_{i=1}^{|M|} u_{il} = 0 \quad l \in B; \quad l \neq P \tag{4.4}$$

For all other nodes k in the network except the target node P the number of in-going arcs must be equal to the number of out-going arcs, as given in constraint (4.5).

$$\sum_{i=1}^{|M|} u_{ik} = \sum_{j=1}^{|M|} u_{kj} \quad k \in M \setminus S; \quad k \neq P \tag{4.5}$$

This means that if an arc is entering an intermediate node k, then there must also be an arc leaving this node. Constraints (4.3) to (4.5) ensure that a path can only start with a start metabolite contained in the set of possible start nodes S. This constraint was taken from (PEY et al., 2011), but has been adapted for start metabolites. Constraint (4.6), which was adopted from (PEY et al., 2011), forces nodes on a path to be unique, i.e. at most one arc can enter any given node.

$$\sum_{i=1}^{|M|} u_{ik} \leq 1, \quad k = 1, \ldots, |M| \tag{4.6}$$

Constraints (4.1) to (4.6) ensure that a solution contains a connected simple path from a start node of the set of start nodes S to a given end node P.

The next set of constraints deals with the feasibility of the linear path in the given network. Given are the stoichiometric coefficients S_{mr} for a metabolite m in reaction r (for $m = 1, \ldots, |M|$, $r = 1, \ldots, |R|$). The variables v_r assign each reaction r a non-negative flux. Constraint (4.7) expresses that the external metabolites are not necessarily balanced and can only be produced, but not be taken up. Only metabolites from the metabolite pool E_m containing the set of start metabolites, basis metabolites, cofactors and excluded metabolites can be taken up. This means that all substrates on the pathway must be producible with metabolites contained in the metabolite pool. This constraint was adopted from (PEY et al., 2011).

$$\sum_{r=1}^{|R|} S_{mr} v_r \geq 0, \quad \forall m \in E, m \notin E_m \tag{4.7}$$

We added constraint (4.8) to make sure the target metabolite P can only be produced.

$$\sum_{r=1}^{|R|} S_{Pr} v_r \geq 1, \tag{4.8}$$

With constraints (4.9) and (4.10), (adopted from (PEY et al., 2011)), we introduce the binary variable z_r which is 1, when reaction r has a flux and 0 otherwise. All fluxes are scaled between 1 and a chosen positive value Max with $Max \geq 1$. This constraint relates fluxes in the flux distribution defined by v_r to reactions.

$$z_r \leq v_r, \qquad r = 1, \ldots, R \tag{4.9}$$

$$\text{and} \quad v_r \leq Max \cdot z_r, \qquad r = 1, \ldots, R \tag{4.10}$$

Constraint (4.11) states that a reaction and its reverse cannot appear together in a valid flux distribution to exclude trivial cycles. This constraint was adopted from (PEY et al., 2011)).

$$z_\lambda + z_\mu \leq 1 \qquad (4.11)$$
$$\forall(\lambda,\mu) \in B = \{(\lambda,\mu)|\lambda \text{ and } \mu \text{ are reverse}\}$$

The path-finding and the stoichiometry constraints are linked through a linking constraint (4.12).

$$\sum_{r=1}^{|R|} d_{ijr} \cdot z_r \geq u_{ij} \quad i = 1,\ldots,|M|; j = 1,\ldots,|M|; i \neq j \qquad (4.12)$$

The binary coefficients d_{ijr} are 1, if there exists an arc between the metabolites i and j in reaction r and 0 otherwise. If an arc from i to j is used in the path ($u_{ij} = 1$) then at least one reaction r containing this arc ($d_{ijr} = 1$) has to be active. This constraint was adopted from (PEY et al., 2011)).

Constraints (4.7) to (4.12) define a valid flux distribution for the pathway ensuring that the found path is feasible.

The objective function of the problem is formulated in equation (4.13).

$$Minimize \quad \sum_{i=1}^{|M|} \sum_{j=1,j\neq i}^{|M|} u_{ij} + \frac{1}{|R|+1} \sum_{i=1}^{|R|} z_i \qquad (4.13)$$

As proposed by (PEY et al., 2011) we minimize the number of arcs u_{ij} used but additionally we also minimize the number of active reactions on the whole pathway candidate.

A solution to the MILP described by equations (4.1) to (4.13) is a sequence of arcs given by the values of u_{ij} and the set of active reactions given by the values of z_r. By minimizing the objective function we ensure that the linear path is connected and cycle-free and the number of active reactions and thus of supplying reactions is minimal. From the active reactions we determine those corresponding to the active arcs, denoted as Z'. One solution represents one pathway candidate.

To find further solutions we have to exclude solutions with the same active arcs and the same reactions Z'. Note that a valid new solution can have exactly the same set of active arcs as a previous solution if Z' is different, since an arc can be derived from more than one reaction. Let U_{ij}^k be the value of u_{ij} for the k-th unique solution with respect to the metabolite path.

To indicate that a solution is exactly the same as solution k regarding the metabolite path, we introduce a binary variable s_k. When a solution is different from solution k regarding the metabolite path, s_k has to be 0 and 1 otherwise. Whenever we find a metabolite path $U^{k'}$ we have not seen before, we introduce constraints (4.14) to (4.16) and a new binary variable $s_{k'}$.

$$\sum_{i}^{|M|}\sum_{j}^{|M|} U_{ij}^{k'} \cdot s_{k'} \leq \sum_{i}^{|M|}\sum_{j}^{|M|} U_{ij}^{k'} u_{ij} \tag{4.14}$$

$$\sum_{i}^{|M|}\sum_{j}^{|M|} (1 - U_{ij}^{k'}) u_{ij} + s_{k'}|M|^2 \leq |M|^2 \tag{4.15}$$

Constraints (4.14) and (4.15) establish that, whenever we find a new solution U and $s_{k'}$ is set to 1, we know that $U = U^{k'}$. In more detail, constraint (4.14) ensures that if $s_{k'}$ is 1 all arcs of solution k' are also active. Additionally, constraint (4.15) forbids U to contain any arc that was not present in $U^{k'}$. We denote the first metabolite in the path in solution k' by $\alpha^{k'}$.

$$\sum_{i}^{|M|}\sum_{j}^{|M|} U_{ij}^{k'} u_{ij} - \sum_{i}^{|M|} u_{i\alpha^{k'}} - s_{k'} \leq \sum_{i}^{|M|}\sum_{j}^{|M|} U_{ij}^{k'} - 1 \tag{4.16}$$

Constraint (4.16) ensures that a valid new solution has to fulfill one of the following three properties. It has either exactly the same metabolite path $U^{k'}$; or at least one of the arcs from the previous metabolite path $U^{k'}$ is not active; or all arcs from $U^{k'}$ are active and one arc entering the first metabolite $\alpha^{k'}$ is active extending a previously found metabolite path. This constraint also ensures that $s_{k'}$ is set to 1 if $U = U^{k'}$. Constraint (4.17) is always added for each new solution. Assume the found metabolite path is the same from solution k (U^k). Let $Z_i'^l$ indicate whether reaction i is active in solution l and corresponds to an active arc in U^k. The number of ones in Z'^l is denoted by m_l. This constraint prevents to find a second solution that is exactly the same as a previously found solution with regard to both linear path and reactions.

$$\sum_{i}^{|R|} Z_i'^l z_i + s_k|R| \leq m_l - 1 + |R| \tag{4.17}$$

Figure 4.3 depicts an exemplary pathway to the target metabolite P illustrating a possible

solution of the presented MILP. The light yellow square **M1** is the starting metabolite of the

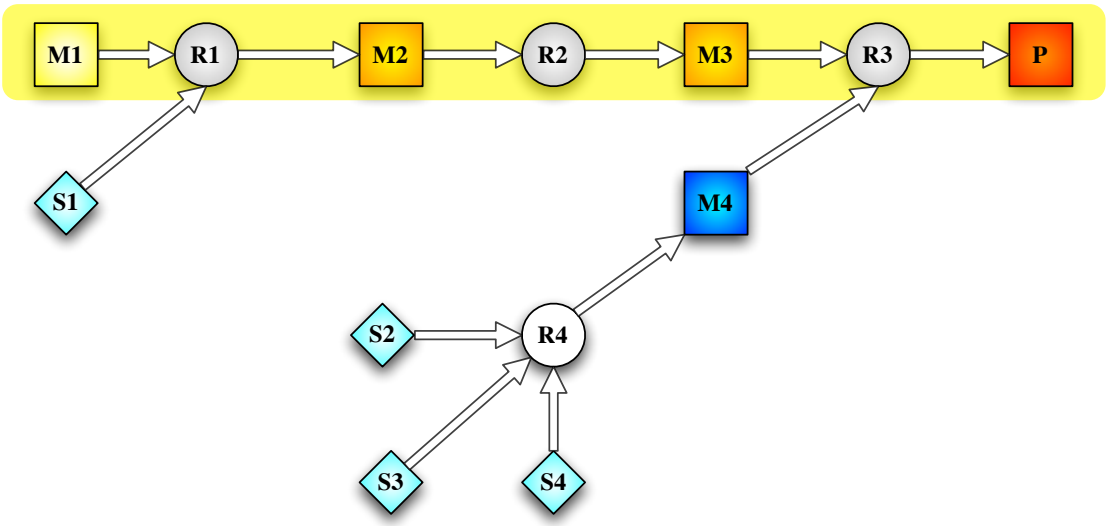


Figure 4.3: Exemplary pathway illustrating a possible solution. The squares depict metabolites, the circles represent reactions. The pathway is a feasible synthesis pathway from M1 to the product P.

linear path, whereas the dark orange square **P** is the target metabolite. The light blue squares are metabolites from the metabolite pool. The linear path highlighted in yellow is defined through constraints (4.1) to (4.6). One of the substrates for reaction **R3**, metabolite **M4**, is not available in the metabolite pool and thus must be supplied by other reactions. These supplying reactions are defined by constraints (4.7) to (4.12). In this example, reaction **R4** depicted by the white circle is added to the resulting path. The overall pathway is a synthesis pathway from **M1** to the desired product **P** that is feasible within the given network.

Filtering and Ranking

We rank the pathway candidates generated by the MILP by different criteria in order to highlight the most meaningful candidates for the synthesis of the desired product. As a global optimization method, the MILP cannot take into account if the first reaction of a pathway candidate is feasible only with metabolites in the metabolite pool. We thus have to perform a filtering step before the ranking to eliminate those pathway candidates that do not comply with this requirement. The ranking criteria are listed in Table 4.1.

The first criterion is the number of active reactions in the pathway candidate. Shorter pathways favor a fast product formation, a reduced substrate demand and are generally

Table 4.1: Ranking criteria in the order they are applied to the pathway candidates.

position	criterion	comment
1	number of active reactions	shorter pathways are favorable
2	candidate starts with basic metabolites only	'yes' is preferred
3	number of reactions without $\Delta_r G$	as few as possible
4	$\sum(\Delta_r G + \lvert\Delta_r G\rvert)$	preferably all $\Delta_r G$ are negative
5	$\sum \Delta_r G$	negative is preferred
6	number of heterologous enzymes	as few as possible
7	number of cofactors	as few as possible

easier to realize than a pathway with more reactions. The second ranking criterion prefers pathway candidates starting with basic metabolites only. A further ranking criterion favors pathways for which there is thermodynamic information available. This is based on the notion that reactions without known or assessable $\Delta_r G$ are often poorly described. Another ranking criterion is the sum of the $\Delta_r G$'s and the absolute value of those $\Delta_r G$'s $\sum_r(\Delta_r G + |\Delta_r G|)$ for all reactions r in the linear path of the pathway candidate. Ideally this sum is 0, since then each reaction has a negative $\Delta_r G$. Therefore, pathway candidates with positive $\Delta_r G$ of intermediate reactions are ranked down, as they would lead to kinetic traps. Furthermore, the pathway candidates are ranked by the overall thermodynamics of the linear path of the pathway candidate. Pathways with a negative overall $\Delta_r G$ are preferred over those with a positive overall $\Delta_r G$. The ranking also takes into account the number of enzymes that are native in a specified host organism. Pathways with less heterologous enzymes are preferred as they potentially require less genetic engineering work in the practical implementation. The last ranking criterion counts the number of different cofactor species that are required by a pathway candidate. Cofactors are often expensive and require regeneration which can be difficult to implement. Thus, pathway candidates with less cofactors are preferred.

In addition to the output of the reactions of each pathway candidate and an overall balance of each reactant in a pathway, further information useful for their assessment is given. The thermodynamic profile allows for a quick visual assessment of each pathway. An SBML (HUCKA et al., 2003) file containing all reactions on the pathway allows the visualization of the path and the active reactions with any tool capable of reading SBML (e.g. Cytoscape (SHANNON et al., 2003; SMOOT et al., 2011)). A list of possible side reactions for each pathway candidate in a given host organism can help to find pathways with a small number of side reactions or even identify those side reactions that can be deleted.

Computational Details

Our path-finding tool is implemented in MATLAB© R2015a (8.5.0) (MathWorks). As a MILP solver we used the IBM CPLEX Optimizer 12.5. All data from KEGG is obtained using the KEGG REST API. The eQuilibrator 2.0 source code was cloned from the GitHub repository of component-contribution [1]. The computations were carried out on a 64bit, 3.4Ghz Intel Core i7-2600 PC with 8 GB RAM.

All data generated or analyzed during this study are included in this part and Appendix A. The software used in this study is available for download at `https://doi.org/10.5281/zenodo.816174`. The most recent version can be found at our github repository `https://github.com/mecatsb/mecat`.

4.3 Results

We use GPP as a first example to illustrate features of our method. GPP is part of the metabolism of most organisms and plays a key role in the terpenoid biosynthesis. Its precursors isopentenyl pyrophosphate (IPP) and dimethylallyl pyrophosphate (DMAPP) can be synthesized via two different pathways. The mevalonate pathway starting with acetyl-CoA is present in fungi, archaea and some bacteria. The non-mevalonate pathway (MEP/DOXP pathway) with pyruvate as a precursor exists in plants, eubacteria and protozoa (MICHAL et al., 2012). From the computed pathways we chose interesting candidates depicted in Figure 4.4 and 4.5.

The pathway candidate in Figure 4.4 corresponds to the lower mevalonate pathway. It starts with 2-oxoglutarate synthesizing IPP and DMAPP in seven consecutive reactions plus an additional reaction to GPP. The pathway candidate has 11 potential side reactions which are provided in more detail in the Appendix Section A.1. These reactions can potentially be active in permeabilized cells or cell lysates but might be disrupted by corresponding gene deletions. If a synthetic mixture of enzymes of interest would be applied, these reactions would not be active at all. With the presented network we were also able to recover the non-mevalonate pathway shown in Figure 4.5. The thermodynamic profiles for the linear path of these pathways are shown in Figure 4.6 and 4.7. They indicate that the operation of these pathways is thermodynamically feasible with negative and constantly dropping $\Delta_r G$. Our tool proposes 11 potential side reactions for the mevalonate pathway and 24 for the

1 `https://github.com/eladnoor/component-contribution`, cloned on 20.02.2016

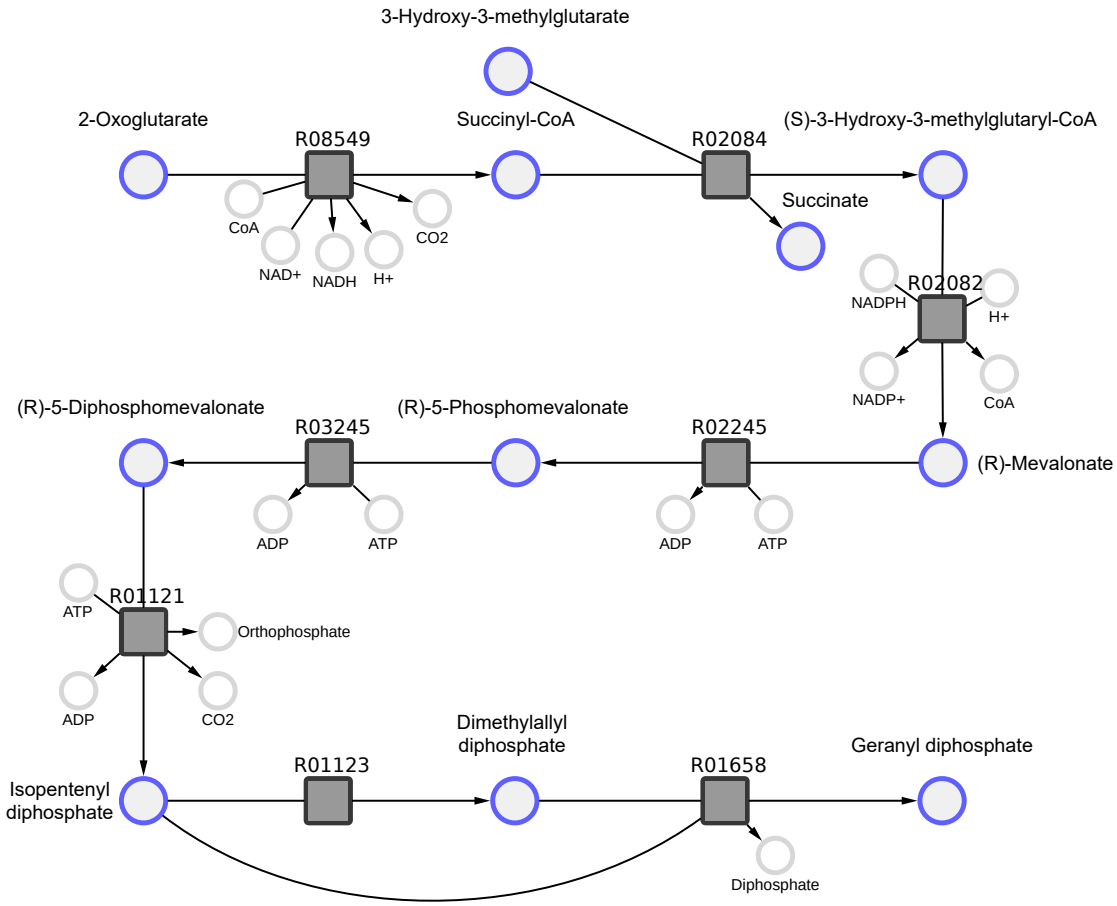


Figure 4.4: Pathway candidate 1. Synthesis of geranyl pyrophosphate via the mevalonate pathway.

non-mevalonate pathway. They are provided in more detail in the Appendix Section A.1. The candidate for the mevalonate pathway was chosen because of its favorable thermodynamic profile (Figure 4.6) with a large drop of $\Delta_r G$ in the last two reactions. This final drop has the potential to lead to high conversion. Additionally, all substrates for the synthesis are readily available. However, the mevalonate pathway is not natively present in our chosen host *E. coli*. The second pathway candidate based on the non-mevalonate pathway displays an alternative method for the production of GPP, which is fully present in *E. coli*.

We chose amygdalin as a further example. In this case, we added sucrose as a potential starting and basis metabolite. Sucrose is excluded from the original set of starting metabolites because of its higher molecular mass but is much cheaper than α-D-glucose 6-phosphate. The generated pathways contain two interesting candidates with both four consecutive

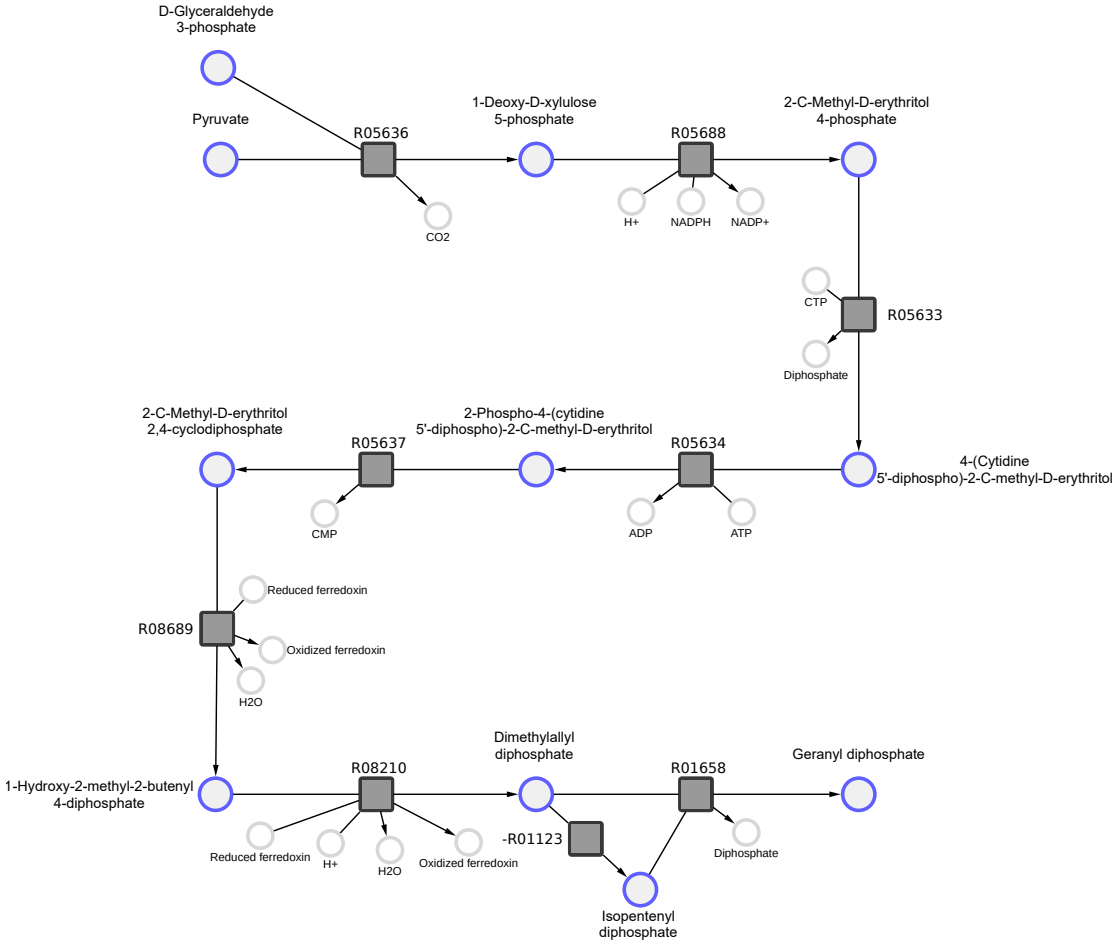


Figure 4.5: Pathway candidate 2. Synthesis of geranyl pyrophosphate via the non-mevalonate pathway.

active reactions to amygdalin. The first candidate starts with sucrose and the second with α-D-glucose 6-phosphate.

Both candidates require a uridyl moiety as substrate. Nevertheless, in the search carried out, uridine 5'-triphosphate (UTP), uridine 5'-diphosphate (UDP) and uridine 5'-monophosphate (UMP) were considered cofactors to avoid unnecessary interconversion of nucleotides that would add numerous but not meaningful pathway candidates. And in both candidates, two of the reactions are catalyzed by heterologous enzymes. For the first pathway, four potential side reactions are proposed and five for the second. These pathway candidates highlight the impact of the list of potential starting metabolites on the results. While both pathways look promising, the first one starts with the cheap starting substrate sucrose and

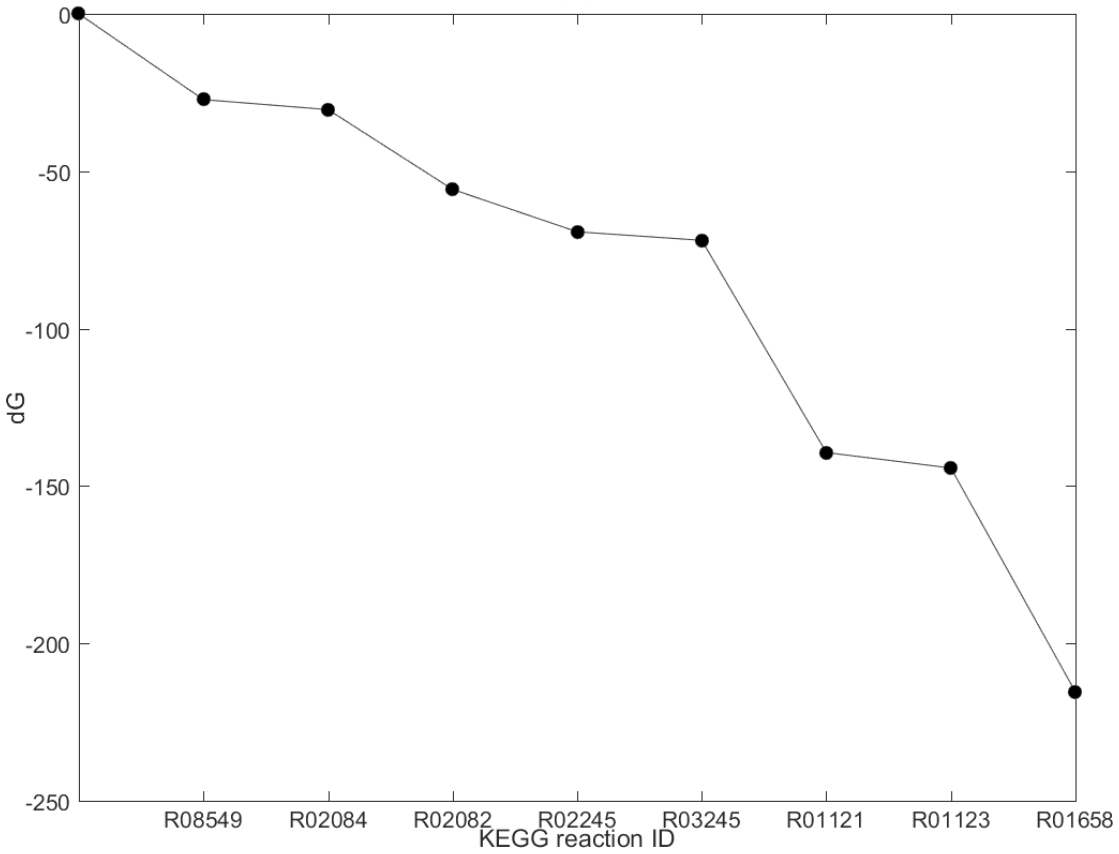


Figure 4.6: Thermodynamic profile for the mevalonate pathway.

has a better thermodynamic profile. In an industrial environment it would be advisable to create a customized list of starting metabolites considering more criteria, e.g. of cost and availability.

Another example is pyrrolysine. The selected pathway candidate has four active reactions and starts with L-Lysine as substrate. Thermodynamic data for this pathway is not available in eQuilibrator. In *E. coli*, this pathway does not exist, but it is native in methanogenic archaea. The pathway requires ATP and NAD^+/reduced nicotinamide adenine dinucleotide (NADH) as cofactors. It has nine potential side reactions.

As a last example, we chose (S)-2-phenyloxirane. The selected pathway candidate for (S)-2-phenyloxirane has four consecutive active reactions. It uses cinnamaldehyde as substrate and requires CoA, $NADP^+$/NADPH and AxP as cofactors. The thermodynamic profile is not ideal with regard to the first and last reaction steps that both have a slightly positive

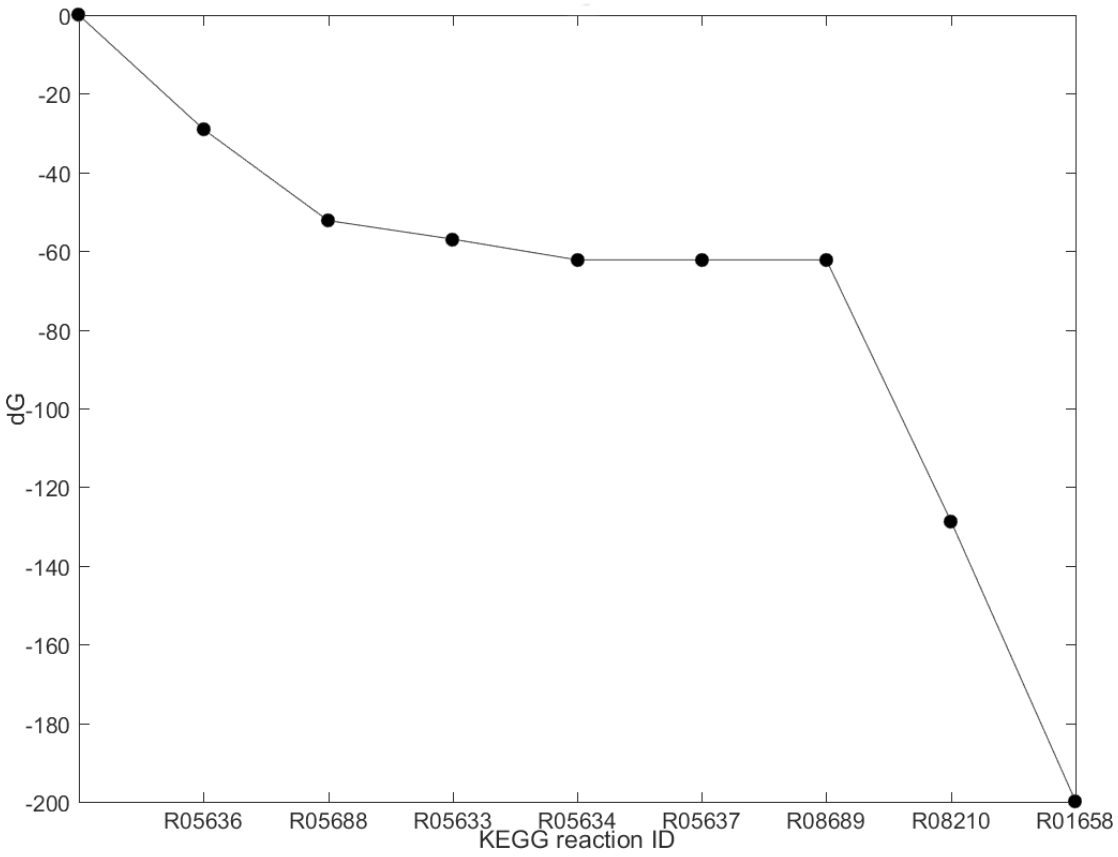


Figure 4.7: Thermodynamic profile for the non-mevalonate pathway.

$\Delta_r G$. Potentially, the last step could be promoted by an efficient reduced flavin adenine dinucleotide ($FADH_2$) regeneration or oxygen supply pushing the equilibrium to the product side. However, it remains questionable if $FADH_2$ can be regenerated in permeabilized cells.

Details to all examples shown are given in the respective sections of the Appendix Section A.1. The Appendix Section A.2 contains details on the computation times of all examples.

4.4 Discussion

We presented a method for searching potential synthesis pathways for target metabolites without the specification of a fixed starting point. Due to the nature of the search algorithm, the resulting pathway candidates are unbiased by the user's knowledge and expectation of the most suitable pathway. Our method leads to a large number of results in a broad solution space which may make it challenging to find the most appropriate candidate. Handling

this amount of data requires a sophisticated tool of filtering, ranking and expert assessment together with additional features such as the quick evaluation of potential side reactions and thermodynamics. Altogether, our tool is highly customizable and offers flexible filtering and ranking options. All metabolite lists, especially the metabolite pool can be easily adapted to meet the needs of a specific project. This is especially useful in cases where the metabolite pool should be composed of chemicals of the laboratories' inventory or of inexpensive chemicals. Analogously, all ranking or filtering criteria can be tailored to the focus of the study, such as reagent costs or a specific host organism.

Expert knowledge to assess the pathway candidates is still needed. However, the same applies to any pathway design method available to date. The resulting pathway candidates depend fully on the data used to set up the network. The sheer mass of reactions in KEGG makes errors hard to identify manually, and we did not carry out any data cleaning except the measures discussed in section 4.2 Network Reconstruction. Crude errors such as unbalanced or ill-formed reaction entries in KEGG were automatically identified and excluded from our network.

Thermodynamics of a pathway is complex. Most substances involved in a pathway are not present at the beginning but are rather formed as the synthesis proceeds. This is not taken into consideration. We fix the initial starting concentrations of all metabolites to 1 mM. However, these can be easily modified by adapting the respective values for the calculation of the $\Delta_r G$ in eQuilibrator. Note, that all $\Delta_r G$ are estimated using the component contribution method. They can however be replaced by experimental values, if available.

We do not consider enzyme concentrations or any kind of kinetic parameters such as enzyme turnover numbers or K_m values. While this would be a relevant addition, to our knowledge this information is not readily available on the scale needed for large networks. It could however be integrated for smaller networks, e. g. (KHODAYARI et al., 2016), particularly in the ranking procedure.

4.5 Conclusions

The presented method provides a helpful computational tool for the directed design of biosynthetic production pathways and the planning of syntheses. The tool provides a very useful basis for the eventual selection of pathways to be implemented in the wet lab. Building on this, expert knowledge is required to tackle possible practical problems with the implementation of the most promising candidates. All features presented are autonomous. The generated

thermodynamic profiles of pathways are invaluable for selecting the most promising pathway alternatives. Similarly, computing potential side reactions leads to important insights for all kinds of pathways.

In different use cases different ranking criteria may be considered important. The user of the tool can easily select or define own criteria for ranking results. For the synthesis with cell lysates or permeabilized cells, the consideration of heterologous enzymes and the choice of the most suitable host as well as potential side reactions are certainly very important.

Part III

Network Reconstructions for Cell-Free Systems

CHAPTER 5

In-depth Characterization of Genome-Scale Network Reconstructions for the *in vitro* Synthesis in Cell-Free Systems

The following chapter is based on the published research article

> SCHUH, L. K., WEYLER, C., & HEINZLE, E. (2019): In-depth characterization of genome-scale network reconstructions for the in vitro synthesis in cell-free systems. *Biotechnology and Bioengineering* (2019), vol. *117*([4]): 1137–1147. https://doi.org/10.1002/bit.27249

Appendix B is based on the the supplementary material of this research article.

Abstract

Cell-free systems containing multiple enzymes are becoming an increasingly interesting tool for one-pot syntheses of biochemical compounds. To extensively explore the enormous wealth of enzymes in the biological space, we present methods for assembling and curing data from databases to apply them for the prediction of pathway candidates for directed enzymatic synthesis. We use KEGG to establish single organism models and a pan-organism model that is combining the available data from all organisms listed there. We introduce a filtering scheme to remove data that are not suitable, e.g. generic metabolites and general reactions. Additionally, a valid stoichiometry of reactions is required for acceptance. The networks created are analyzed by graph theoretical methods to identify a set of metabolites that are potentially reachable from a defined set of starting metabolites. Thus, metabolites

not connected to such starting metabolites cannot be produced unless new starting metabolites or reactions are introduced. The network models also comprise stoichiometric and thermodynamic data that allow the definition of constraints to identify potential pathways. The resulting data can be directly applied using existing or future pathway finding tools.

5.1 Introduction

The enzymatic potential of the numerous enzymes in nature is a most promising, extremely versatile and powerful resource for creating powerful tools for the production of various interesting products. Besides the production in host organisms, synthesis using cell-free systems gains more and more interest. Particularly multi-step biocatalysis seems only marginally explored today compared to its expected huge potential (HEINZLE et al., 2013). Cell-free systems for the synthesis range from mixtures of isolated enzymes over multi-enzyme systems, e.g. multi-enzyme complexes (S.-Z. WANG et al., 2017) and enzyme cascades, to cell lysates (ENDO et al., 2001) and permeabilized cells. In special cases such systems are even combined with chemical synthesis in one pot (GROEGER et al., 2014).

The design of a multi-step synthesis route does not only require the determination of the reaction sequence leading to the desired product, but also depends on numerous aspects such as substrate and cofactor supply or thermodynamics. For living cells, a recent review article discusses the state of the art computational tools for design and reconstruction of metabolic pathways (L. WANG et al., 2017). To design such a pathway for cell-free biosynthesis is by far not developed to such a mature state. In particular, it seems almost impossible to explore manually all potentially feasible pathways and to determine which one is the most suitable for production.

The *in silico* path-finding and design methods all require a metabolic network model containing all required information from the host organisms of interest, such as enzyme, reaction and thermodynamics data. There is an ever-growing plethora of biological databases with enzyme and reaction data of an ever-growing number of organisms that is suited for the reconstruction of genome scale metabolic networks. One of the most popular databases is KEGG (KANEHISA, FURUMICHI, et al., 2016; KANEHISA et al., 2000; KANEHISA et al., 2018). However, despite the huge amount of data collected from primary literature that is carefully curated afterwards, the data is partly incomplete, and sometimes even inconsistent or erroneous. It is thus a challenge to handle this data and make it suitable for useful network reconstructions.

We already presented a computational tool to guide and support finding the most suitable synthesis path to a product (BLASS et al., 2017). We extended this work by developing a method of building network models from KEGG data which is suitable for path-finding. We selected nine organism networks that are of interest primarily for their application in cell-free production. Some were selected because of peculiarities of the networks. Finally, a so-called pan-organism network was used lumping all metabolic reactions listed in KEGG in one single network.

5.2 Materials and Methods

In the following we give a short introduction to our path-finding method. We also present how to build network reconstruction models based on data found in biological databases, particularly KEGG.

Path-finding

We already presented a method for finding candidates for suitable synthesis pathways in genome-scale metabolic network reconstructions starting from arbitrary substrates (BLASS et al., 2017). A *pathway* in our definition consists of two parts. First, the so called *linear path* consists of a sequence of metabolites connected by reactions. It starts with a reaction that has one of the possible predefined start metabolites as a substrate and ends with a reaction that has the target metabolite T as a product. Second, there is the set of *supplying reactions*, which provide the substrates required by the reactions on the pathway that are not contained in the metabolite pool. All metabolites in this pool are considered freely available since they will be provided by the specified pathway reactions (see Section 5.2 Model Building).

The path-finding algorithm is based on a MILP and combines graph-based path-finding and reaction stoichiometry (PEY et al., 2011). The method is elaborated in detail in (BLASS et al., 2017). Figure 5.1 shows an exemplary pathway illustrating a possible solution of the MILP. The pathway shown is a feasible synthesis pathway to the target T (depicted as red octagon). Metabolites in the figure are depicted as squares, where large squares represent metabolites in arcs (see 5.2 Model Building) and small squares represent cofactors and inorganic metabolites. Reactions are represented by circles. The linear path of the pathway is marked with a blue background. Metabolites S1 to S4 and M1 (marked in green) are contained in the metabolite pool (see 5.2 Model Building) and are thus initially available. As M4, which is required by reaction R3, is not available from the metabolite pool, R4 is needed as a supplying reaction producing it.

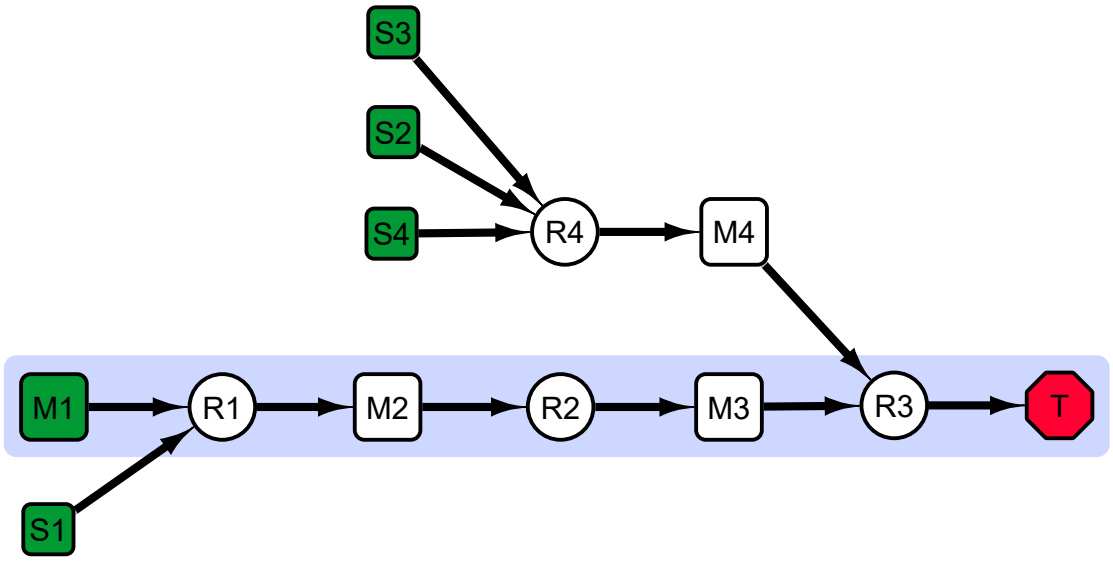


Figure 5.1: Exemplary pathway illustrating a feasible pathway to the target metabolite T (red octagon). Large squares: metabolites with arcs (see 5.2 Model Building); small squares: cofactors/inorganic metabolites; green: metabolites from the metabolite pool (see 5.2 Model Building); circles: reactions; blue background: linear path; R4 : supplying reaction.

In addition to the 17 constraints of the MILP presented in (BLASS et al., 2017) we added a constraint which prevents the use of a reaction in the pathway (more precisely, the supplying reactions) that consumes the target. This constraint is necessary to prevent cycles formed by a reaction belonging to the linear path that produces the target and a supplying reaction consuming the target to produce a precursor which is consumed by a reaction on the linear path. It thus prevents pathways for which the target has to be already present in at least catalytic amounts. An example for such an undesired pathway is shown in Figure 5.2. In this example, the target T needs to be consumed by reaction R6 to form metabolite M4 which is required by reaction R5 to produce the target T.

The complete MILP is listed in Appendix Chapter B.1.

Model Building

In the following, we define the different parts of our network reconstruction and model based on KEGG data. The reactions and metabolites in the model are given as lists of KEGG REACTION and COMPOUND ids (KANEHISA, FURUMICHI, et al., 2016). The reactions and metabolites are connected by arcs, which are derived from reactions.

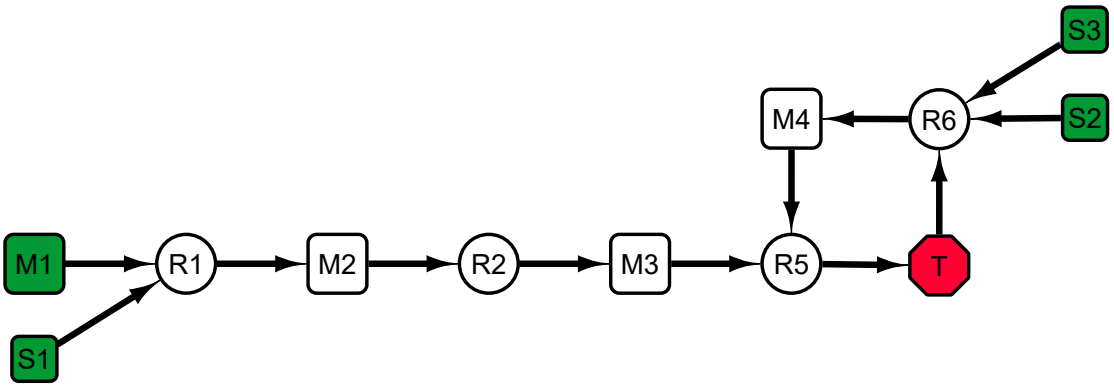


Figure 5.2: Exemplary pathway illustrating a pathway to the target metabolite T (red octagon) where T needs to be consumed in order to produce M4. Large squares: metabolites with arcs; small squares: cofactors/inorganic metabolites; green: metabolites from the metabolite pool; circles: reactions; blue background: linear path; R4 is a supplying reaction. This pathway example is not a valid synthesis pathway candidate for T.

The metabolites in the model are categorized into sets that are treated differently in the path-finding algorithm. One set consists of potential *start metabolites*. These are all metabolites in the model that can be used as the start of the linear path of a pathway candidate. Metabolites in this category are automatically determined and have a molecular mass smaller than 300 and occur in arcs. The so called *basis metabolites* are expert-curated metabolites which are inexpensive, easily available and are often hubs in the arc network, such as D-glucose (C00031) or pyruvate (C00022). The *cofactors* (e.g. ATP (C00002), NADH (C00004) etc.) and *inorganics* such as water (C00001), oxygen (C00007) or CO_2 (C00011) are a set of expert-curated metabolites that are considered as freely available if they are required as substrates in reactions, but are not part of the reaction chain. They are thus excluded from the arcs to prevent biologically meaningless shortcuts in the pathways. All metabolite sets are disjoint, except for the basis metabolites that form a subset of the start metabolites. The *metabolite pool* is the superset of metabolites that are considered as freely available. It is made up of start metabolites, basis metabolites, cofactors and inorganic metabolites. Further details on the different categories are given in Section 4.2 Network Reconstruction.

For each reaction there is a set of arcs, which are substrate-product pairs of a reaction. There are different strategies to derive the arcs from a reaction. The straightforward method is using all possible combinations (i.e. the cross product) of substrates and products of a reaction. It is however more useful to use meaningful substrate-product pairs, such as reactant pairs. A reactant pair is a substrate-product pair with both parts having atoms or atom groups in common that preserves the chemical substructures of the reactants through

the reaction (KOTERA, HATTORI, et al., 2004; KOTERA, OKUNO, et al., 2004). The reactant pairs are defined in the KEGG RCLASS database, which classifies reactions based on the chemical structure patterns of their substrate-product pairs (MUTO et al., 2013). Only those reactant pairs are used for the arcs that do not contain any metabolite from the cofactor and inorganics list. This means, however, that reactions involving metabolites from this list are still represented by the remaining arcs. A more detailed discussion on the arcs can be found in Appendix B.2 (Tables B.12 to B.16). The arc graph of the model is a directed graph G = (V, E), where V is the set of metabolites and E is the set of arcs between these metabolites. The model also contains a stoichiometric matrix, where each row corresponds to a metabolite in the model and each column indicates a reaction. An entry in the matrix is the stoichiometric coefficient of the metabolite in the respective reaction.

When using KEGG COMPOUND and KEGG REACTION data for a network reconstruction some obstacles have to be addressed. One of them is reaction directionality. For the reactions contained in KEGG the reaction directions are not indicated in the database entries. There is thus a need for further reaction data to annotate directionality. To do so, we use the component contribution method of the biochemical thermodynamics calculator eQuilibrator (FLAMHOLZ et al., 2012; NOOR et al., 2013) to compute the $\Delta_r G'^m$ value (the change of the Gibbs free energy of a reaction at a given pH of 7 and ionic strength I in 1 mM concentration of the reactants) for each reaction in the network and infer if the respective reaction is reversible. Reactions with $|\Delta_r G| \leq 15$ kJ/mol are designated as reversible. In biological systems as well as in most biosynthetic setups concentrations of substrates and products often differ by several orders magnitude. This significantly influences reaction reversibility. As these effects cannot be adequately considered given the size of the networks presented in this work and the unknown kinetics, the $\Delta_r G$ value of 15 kJ/mol was chosen as a consensus value to determine reaction reversibility. This somewhat arbitrary value represents a compromise between the assumption of reversibility of all reactions and a more stringent restriction with a $\Delta_r G$ value of less than 15 kJ/mol that would potentially exclude feasible biosynthetic routes with concentrations of intermediates adjusting in a running system. The value was set after a series of simulations and expert inspection of results. However, the user of our tool can freely set the $\Delta_r G$ cutoff to meet the needs of his specific investigation. The reactions are added to the model in the respective direction(s), which means that for each reversible reaction we get two reactions in the respective directions. Another obstacle is the inconsistent use of identifiers for metabolites. In some reaction equations, the KEGG COMPOUND (C) identifiers are used and in others the G identifiers from the KEGG GLYCAN structure database. As we do not consider glycans, those reactions are excluded. For polymerization reactions, the

reaction stoichiometry in KEGG is not expressed in distinct numbers. Such reactions are not applicable for our method where the coefficients in the stoichiometric matrix are required to be integer numbers.

We did not generally exclude membrane associated reactions. To our knowledge it is not sufficiently clear whether and to which extent intracellular as well as extracellular membrane associated enzymes are active in permeabilized cells. In earlier work we could, however, experimentally show that megasynthases producing a circular oligopeptide can be kept active in permeabilized cells in contrast to cell extracts where activities could not be detected (WEYLER et al., 2017). The exact reasons were not identified but could potentially be related to yet unknown membrane association. On the other hand, in selectively permeabilized eukaryotes, the organelles including membrane reactions remain intact and functional (e.g. (NICOLAE et al., 2015)).

We thus have to filter the KEGG data before building a model. Figure 5.3 shows the filtering steps to obtain the reactions suitable for building a reconstruction of a pan-organism network encompassing reactions from all organisms and also for organism-specific networks.

The filtering starts with all 11196 reactions in KEGG REACTION. First, the reactions with invalid reactants are removed, which are reactants that do not have a C identifier. The 10764 remaining reactions are further trimmed down to 10603 reactions with valid stoichiometry, where all reactants have integer stoichiometric coefficients. From these, reactions that are generic or contain generic reactants (i.e. the database entry has a comment containing 'generic', 'incomplete' or 'general') are removed, sparing 7989 reactions. After removing those without any reaction class annotations, 7676 reactions remain in the pan-organism model, which corresponds to about 69 % of all KEGG reactions.

To build the organism-specific models, the organism annotation for the genes of the enzymes catalyzing those reactions is used. From the 7676 reactions in the pan-organism network reconstruction, KEGG has EC numbers associated with 5975 reactions. 4549 (76 %) of them have enzymes whose genes are annotated with organisms. These reactions are the basis of the organism-specific network model reconstructions. Our network reconstruction workflow filters out ill-formed reaction entries in KEGG. However, we do not include a gap filling step. This would require large manual efforts that are not in the scope of this work.

Possible target metabolites in KEGG for the computation of synthesis pathway candidates are determined automatically. A target metabolite is a metabolite in the respective model that is not a dedicated start or basis metabolite. It also has to appear as a product in at least one

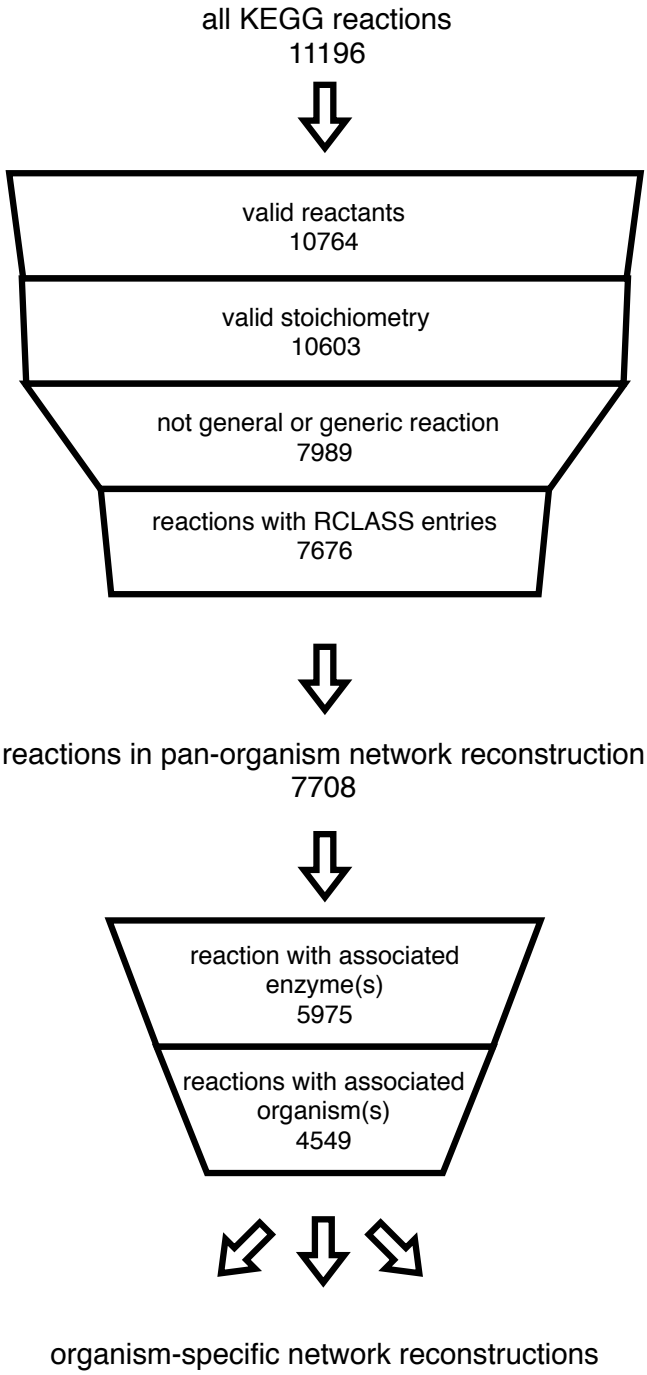


Figure 5.3: Reaction filtering from all reactions in KEGG to the set of reactions for building the pan-organism network reconstruction and the organism-specific models. The reactions are filtered in the given order. The numbers indicate how many reactions stay after filtering. The width of the box bases are proportional to the number of reaction that remain after filtering.

arc in the network, so it could be potentially produced. We predict potentially producible targets in a given model by determining its feasible reactions, i.e. reactions for which potentially all substrates are available or producible. The feasible reactions are obtained by initially starting with the set of metabolites consisting of the model's start metabolites and cofactors/inorganics. With these metabolites, all reactions that are feasible are determined by checking for each reaction that has not already been added to the set of feasible reactions if all substrates are available. The products of these feasible reactions are added to the set of metabolites. This step is repeated until no new substrates are added. The resulting set of reactions is then a subset of the model's reactions that potentially are feasible.

The next step is to do a reachability screening in the arc graph of the model. To do so, we add a node representing an artificial start metabolite that is connected to all potential start metabolites. From there, we do a breadth-first search (BFS), which is a suitable algorithm for exploring a graph. The search starts with a source vertex and discovers all neighboring vertices with the present depth before discovering the next depth-level vertices (CORMEN et al., 2009). The potentially producible targets are those targets that are connected with the start node by a path (a sequence of edges that connect vertices) and that are produced by any of the feasible reactions.

Computational Details

The model data is based on KEGG release 90.1, May 1, 2019. The code for model building and statistics is written in Python 2.7, the code for the thermodynamics is written in Python 3.6 using the eQuilibrator API (FLAMHOLZ et al., 2012). We furthermore used the packages graph-tool (PEIXOTO, 2014) and Matplotlib (HUNTER, 2007). The path-finding tool was run on MATLAB R2019a with IBM CPLEX Studio 12.9. All computations were carried out on an Intel Core i7 with 2.5 Ghz and 32 GB RAM.

The software used in this study is available at `https://github.com/mecatsb`, where the repository `mecat` contains the path-finding tool and the repository `mecatpy` contains the code used for the pathway analysis as well as the organism models. Release `v1.0` contains the code version used in this study.

5.3 Results

We first present the organisms and models used in our study and then discuss some interesting properties of these models. We finally present and discuss the results of our path-finding

analysis.

Models

For each organism in the KEGG Organisms database we build an organism-specific network model as described in Section 5.2 Model Building. Figure 5.4 shows the number of reactions in KEGG that are annotated for the specific organism together with the number of reactions that are part of the organism-specific network reconstruction. The organisms are sorted

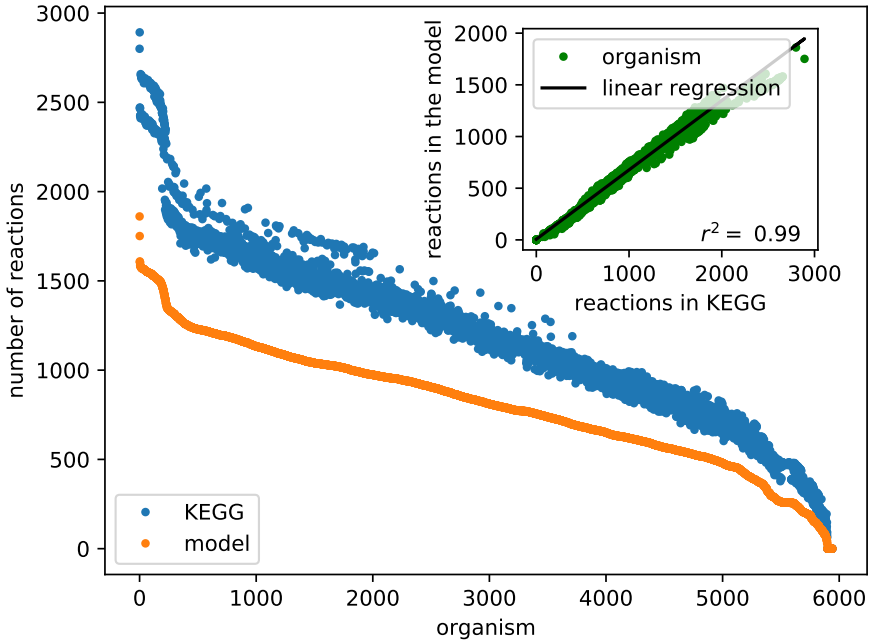


Figure 5.4: Comparison of the total number of reactions and the number of reactions selected for the models for all organisms annotated in KEGG.

in descending order with respect to the number of annotated reactions in the model. The order in which the reactions are filtered is following the procedure shown in Figure 5.3. On average 67 % of the reactions in KEGG that are annotated with an organism end up in an organism-specific model (see insert in Figure 5.4). The reason for this is the filtering of all reactions according to the filter constraints shown in Figure 5.3 and discussed in detail in Section 5.2 Model Building. Figure 5.3 shows that the majority of the discarded reactions are general and/or generic or contain generic reactants.

In addition to the pan-organism network model we chose nine organism-specific models for all network and pathway analyses as examples. Table 5.1 lists the organisms, which were chosen primarily for their importance in biotechnological production as well as in scientific research. CHO, the permanent cells of the ovary of a Chinese hamster *C. griseus*

Table 5.1: Models for the studies. The model names are derived from the KEGG organism codes, except for the pan-organism network model which is named *kegg*. The number of reactions in KEGG refers to the number of reactions that are annotated for the respective organism. The number of reversible reactions is the corresponding subset of the reactions in the model. The feasible reactions are determined as described in Section 5.2 Model Building based on the set of basis metabolites as start metabolites. The basis metabolites are selected as described in Section 5.2 Model Building.

model	name	reactions (KEGG / model / reversible)	feasible reactions	metabolites (model / basis)
kegg	Pan-organism network model	11196/7676/2934	5467	6473/39
cge	*Cricetulus griseus* (Chinese hamster)	2616/1555/639	922	1485/39
eco	*Escherichia coli* K-12 MG1655	1775/1225/511	950	1191/39
vna	*Vibrio natriegens*	1690/1193/489	888	1180/39
ppun	*Pseudomonas putida* NBRC 14164	1683/1199/469	766	1240/37
mxa	*Myxococcus xanthus*	1492/1021/428	684	1077/36
sce	*Saccharomyces cerevisiae* (budding yeast)	1543/1020/403	655	1031/39
spo	*Schizosaccharomyces pombe* (fission yeast)	1408/905/378	592	915/39
cgb	*Corynebacterium glutamicum* ATCC 13032 (Bielefeld)	1122/792/318	534	845/38
mpe	*Mycoplasma penetrans*	371/236/101	166	314/22

were originally isolated already in 1957. CHO is serving as a model cell line for metabolic studies. Most importantly, however, it is most frequently used for the industrial heterologous production of therapeutic proteins (LALONDE et al., 2017; WURM, 2004). The application of animal cells for biosynthetic purposes is easier starting from cell lines like CHO rather

than cells from primary tissues. *E. coli* is probably the most important model organism and is used in all kinds of areas spanning from basic molecular biological work to industrial applications (PONTRELLI et al., 2018). *V. natriegens* is an extremely fast growing marine bacterium that recently got increasing interest. Due to its duplication time of ten minutes it has been in the focus of molecular biology research, e.g. for protein production also in cell-free systems (FAILMEZGER et al., 2018; HOFFART et al., 2017). *P. putida* is known for its diverse biodegradation and biosynthetic capabilities (LOESCHCKE et al., 2015; NIKEL et al., 2016; POBLETE-CASTRO et al., 2012). *M. xanthus* is a model organism for studying social behavior of bacteria with extended signaling networks and secondary metabolite production (WRÓTNIAK-DRZEWIECKA et al., 2016). *S. cerevisiae* is probably the most important eukaryotic model microorganism used very widely and already for a long time for the production of ethanol in alcoholic beverages and biofuel. It is also widely discussed for the production of other metabolites and its broad application is supported by a large toolbox for metabolic engineering (KRIVORUCHKO et al., 2015; NIELSEN, 2019; STEENSELS et al., 2014). The yeast *S. pombe* is a model organism primarily used in molecular and cell biology but is recently also discussed as promising candidate for the expression and secretion of heterologous proteins (TAKEGAWA et al., 2009). *C. glutamicum* is a most important microorganism in the industrial scale production of amino acids but also other metabolic products (BECKER et al., 2016). While these organisms have been used in a vast range of production processes they are also well understood and we assume that KEGG data on these organisms is relatively complete and accurate. *M. penetrans* has the smallest genome of known organisms and its metabolism is very limited (SASAKI et al., 2002). From the present view, the most important organisms for cell-free synthesis are *E. coli*, *S. cerevisiae*, *P. putida* and *M. xanthus*. All model data is part of the GitHub repository `https://github.com/mecatsb/mecatpy`. We exclude plants and algae from our species models since they seem less applicable from the present view on cell-free biocatalysis.

Table 5.2 shows the number of potential targets for the respective model as defined in Section 5.2 Model Building. The arc reachable targets are those targets that are connected to a basis metabolite via an arc path, which is determined by BFS. The feasible targets are targets that are products of feasible reactions as described in Section 5.2 Model Building. The set of potentially producible targets is the intersection of the targets that are connected to a basis metabolite via an arc path and the targets that are products of the feasible reactions.
Table 5.2 shows that a large portion of potential targets is not connected to any of the basis metabolites in the model. For all models, about 32% (in the *S. pombe* model *spo*) to 43% (in the pan-organism model *kegg*) of all potential targets are potentially producible targets. This

Table 5.2: Number of potential targets for each organism model based on basis metabolites as possible start metabolites. Arc reachable targets: targets that are connected to a basis metabolite via an arc path; feasible targets: targets that are products of feasible reactions as described in Section 5.2 Model Building; potentially producible targets: targets that are connected to a basis metabolite via an arc path and that are products of feasible reactions and are thus realistic targets, intersection of the former two columns of the table; % of potential targets: percentage of potentially producible targets in relation to the total number of potential targets.

model	potential targets	arc reachable targets	feasible targets	potentially producible targets	% of potential targets
kegg	5441	3017	2412	2325	43%
cge	1128	437	358	333	30%
eco	878	419	376	351	40%
vna	865	397	348	328	38%
ppun	902	380	317	293	32%
mxa	777	320	281	266	34%
sce	713	268	243	227	32%
spo	637	264	216	201	32%
cgb	598	261	215	200	33%
mpe	184	70	69	56	30%

means that for all other potential targets a synthesis pathway cannot be found, as a path is a required part of a valid solution. We will elaborate the reasons for this drastic reduction later in this work.

Network Model Analysis

We first present some basic properties of the arc graphs of the different organism models. Figure B.1 in Appendix B.2 shows the node degree distributions of the arc graphs of the different organism network reconstructions. The *degree* of a node is the number of edges leaving it (out-degree) plus the number of edges entering it (in-degree). Tables B.1 to B.10 in Appendix B.2 list the hubs with the top 5 occurrences of each network. As expected, pyruvate, L – glutamate, D – glyceradehyde 3-phosphate and acetyl-CoA are in almost all cases metabolites with highest node degrees. *M. penetrans* (*mpe*), having the smallest network of all studied here, differs most significantly from all others both in the types of metabolites with highest node degrees as well as in the generally small numbers of node degrees (< 13). In the pan-organism network model (*kegg*), trans,trans-farnesyl diphosphate has an exceptionally

high node degree (107) that is, however, mostly originating from plant metabolism. In *kegg*, pyruvate is by far the most connected metabolite with a node degree of 167. The outstanding role of only a few metabolites is most strikingly seen in Figure B.1 of Appendix B.2. The sizes of the arc graphs together with the average node degrees, standard deviation of the distribution are listed in Table B.11 of the Appendix Section B.2. It is interesting to see that the average node degrees vary only from 2.37 to 3.14 for individual organisms and 3.3 for *kegg*, the pan-organism network.

Table 5.3 lists the number of connected components of the arc graphs in the respective models and the size of the largest connected component, respectively. A connected component in

Table 5.3: Number of components in the models with the number of metabolites in the largest component. The fourth column lists the number of components containing basis metabolites. The last column shows the number of metabolites that belong to a component containing basis metabolites. The percentage of those metabolites in relation to the number of metabolites in the arc graph is shown in parentheses.

model	number of components	size of largest component	components with basis metabolites	metabolites in components with basis metabolites
kegg	481	4612	1	4612 (74%)
cge	186	754	4	763 (55%)
eco	139	766	2	768 (69%)
vna	146	726	1	726 (66%)
ppun	157	761	2	763 (66%)
mxa	159	587	3	595 (60%)
sce	182	478	1	478 (50%)
spo	162	438	1	438 (52%)
cgb	115	457	2	459 (60%)
mpe	57	76	7	141 (54%)

the graph is a subgraph where each vertex in the subgraph is connected to each other vertex in the subgraph by a path (CORMEN et al., 2009). The smallest connected components contain 2 vertices in all models. This is by definition the smallest component size as the arc graph does not contain metabolites without any arcs. We furthermore list the number of components containing basis metabolites; as well as the total number of metabolites in all those components with the percentage of those metabolites in relation to the number of metabolites in the arc graph (in parentheses). These numbers give information on how much of each network is possibly reachable from the designated start points, since a potentially

producible target has to be connected to any of the predefined basis metabolites via an arc path. Table 5.3 shows that between half and two third of the metabolites in a model's arc graph are contained in a component with basis metabolites. Exemplarily, Figure 5.5 shows the arc graph of the pan-organism model *kegg*. The arc graph consists of a large main

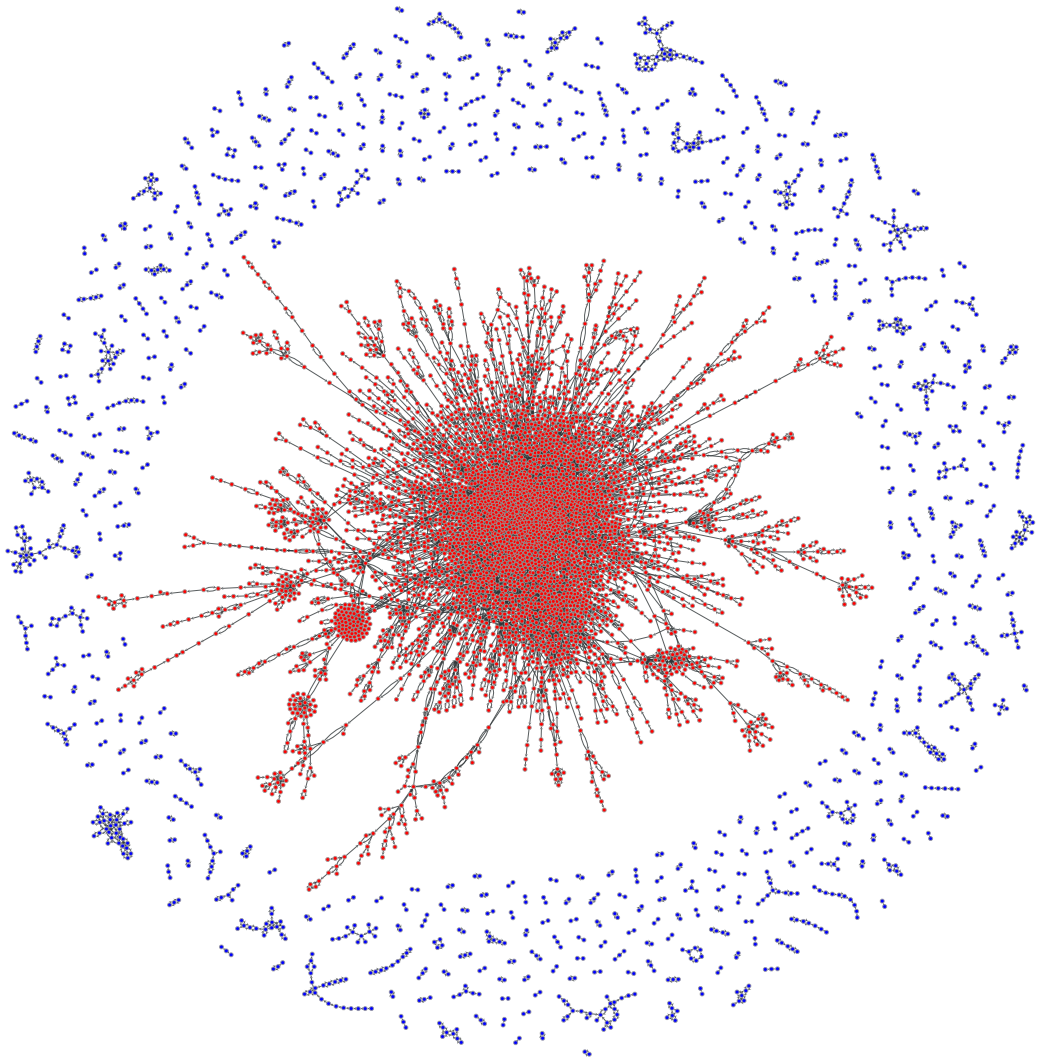


Figure 5.5: Arc graph of the pan-organism model kegg. Red: components containing potential start metabolites; blue: satellite components without start metabolites.

component and a large number of small components. Components in red are components containing potential start metabolites, whereas the components in blue are so called satellite components without start metabolites. The arc graphs of the other models are shown in

Figure B.2 of Appendix B.2. Figure B.3 of Appendix B.2 shows the arc graph component histograms.

There are several reasons for isolated components in a model. The first reason is missing annotation in the data on which the model is based. This could be improved by using manually compiled and curated network reconstructions with gap filling. Several reactions in KEGG are formulated as general reactions and/or are reactions containing generic compounds. Some of the often numerous reactions summarized in such reactions are explicitly listed in KEGG. An even larger number could in principle be added e.g. from BRENDA (JESKE et al., 2018). Some reactions involve additional proteins that transfer electrons or groups or use covalently bound cofactors as e.g. NAD(P)H. These are filtered out in the model building process. As we do not include such reactions in our model, some metabolic pathways could be cut off. Another reason is that a component is really isolated.

In the Supplementary Information 3 of (SCHUH et al., 2019) we list all components identified in the *kegg* model. The 6246 metabolites are grouped in 481 components. The largest component connected to start metabolites comprises 4612 metabolites and is represented in the center of Figure 5.5. All other components with a size of 5 or more metabolites were investigated in more detail (Supplementary Information 4 of (SCHUH et al., 2019)). They comprise 682 metabolites in 69 components. We could identify some typical families related to biochemical characteristics (Appendix B.3). Reactions of xenobiotic compounds, e.g. drugs, were most prominent with 14 components with 141 metabolites followed by polyketides (10/109), carbohydrate derived metabolites (10/86), terpenoids (9/112), compounds with gonane tape nucleus (7/74), fatty acid and lipids related compounds (7/59) and flavonoids (4/54). Xenobiotics are inherently not listed in the starting metabolites. Some of these families have often general reactions or involve generic metabolites, e.g. metabolites contain a group -R that is not explicitly specified. R is later cleaved off the metabolite. Smaller components (< 5) were not analyzed in detail but could often serve as missing links in larger pathways once the connecting reactions could be defined following the criteria specified in 5.2 Model Building.

Reachability Analysis

We determined the target reachability in the pan-organism network and the organism-specific networks by testing the existence of a pathway candidate to each possible target starting with basis metabolites using our MILP presented in Section 5.2 Path-finding. Figure 5.6 shows for each model the percentage of targets for which a synthesis pathway candidate has been

identified and for which a pathway candidate is not accessible and why. The raw data for

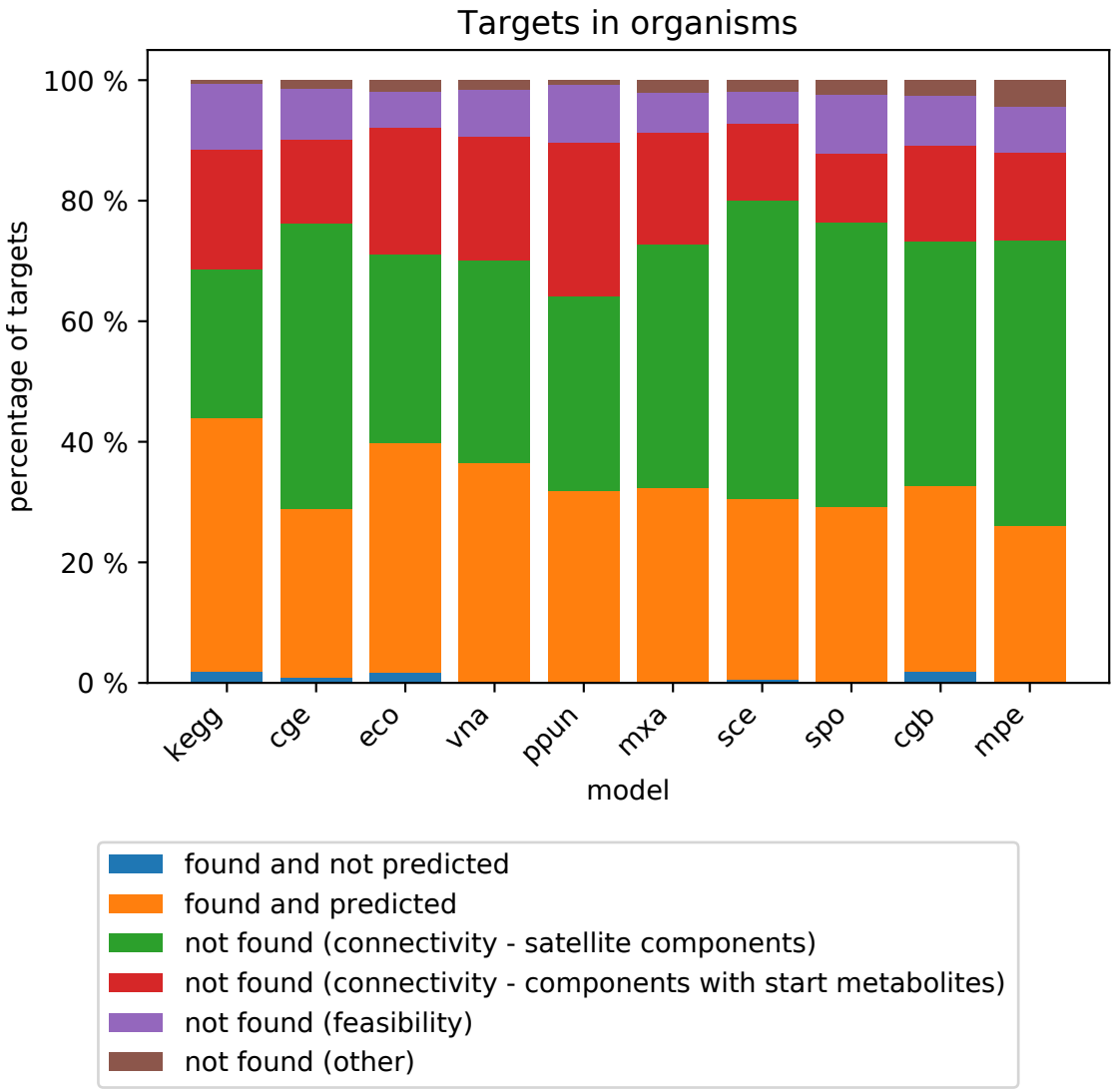


Figure 5.6: Analysis of the target search in the different organism models. Blue: targets for which a pathway candidate has been found by our method, but that have not been predicted as feasible; orange: targets for which a pathway candidate has been found by our method; green: targets for which a candidate has not been found due to the absence of an arc path from any start metabolite to the target because the target is in a satellite component without start metabolite; red: targets for which a candidate has not been found due to the absence of an arc path from any start metabolite to the target (and the target is in a component with start metabolites); purple: targets for which a candidate has not been found due to the lack of a feasible reaction that produces the target; brown: targets for which a candidate has not been found due to other reasons that are discussed in the text.

the figure is listed in Table B.17 of Appendix B.2. In the following, we discuss the different

fractions in more detail.

The blue and orange fractions represent the targets for which a synthesis pathway candidate has been identified in the respective models. The targets represented by the orange fractions have been predicted to have a pathway candidate. This means that they can be produced by feasible reactions of the model and they are connected to at least one of the predefined basis metabolites by a path in the arc graph (see Section 5.2 Model Building). An example for this category is UDP-glucose (Chapter 6 Synthesis Paths for UDP-glucose). However, the targets represented by the blue fractions have not been predicted to be feasible, despite having a synthesis pathway candidate. For those targets we found that most of the pathway candidates calculated with the MILP include a direct cycle formed by supplying reactions that use metabolites that are not in the metabolite pool. In a mathematical sense, it is valid to consume a metabolite as long as its overall balance is zero. However, in real world applications, this would not be correct since the metabolite has to be present in at least catalytic amounts already at the start of the reaction. An example for such a pathway is the pathway candidate for the 5-methyl-5,6,7,8-tetrahydromethanopterin (C04488) production in the pan-organism network model *kegg* (Appendix B.4). The pathway requires coenzyme F420 (C00876) and reduced coenzyme F420 (C01080), which are neither metabolites nor cofactors and thus are not part of the metabolite pool. They thus have to be produced by the reactions of the pathway.

The green, red, purple and brown fractions represent targets without any pathway candidate. In the following, we will discuss the different reasons for this. With the help of BFS, we found that the targets represented by the green and red fractions are not connected to any of the potential start metabolites via an arc path. Therefore, these targets cannot have a pathway candidate, since a path from a start metabolite to the target is mandatory, as stated in Section 5.2 Path-finding.

The targets belonging to the green category are not part of a component containing potential start metabolites. In our pan-organism model *kegg*, this is the case for proansamycin X. Component 134 in Supplementary Information 4 of (SCHUH et al., 2019) shows that there is no reaction in KEGG producing proansamycin X (C12176) from 3-amino-5-hydroxybenzoate (C12107), which belongs to a component with start metabolites (Supplementary Information 3 of (SCHUH et al., 2019)). The situation could be improved by using manually compiled and curated network reconstructions with gap filling, e.g. for metabolites of the earlier discussed polyketide, flavone and terpenoid families (see also Appendix B.3). As outlined in Section 5.2 Model Building, we only did some minor generic curation which has the purpose

of extracting meaningful data and removing ill-specified data. A comprehensive network reconstruction for an organism would require a lot of manual work encompassing more data sources including primary literature, which was not in the scope for this study. However, when using our path-finding method, the user can choose any network model that contains the information needed for path-finding, regardless of data origin.

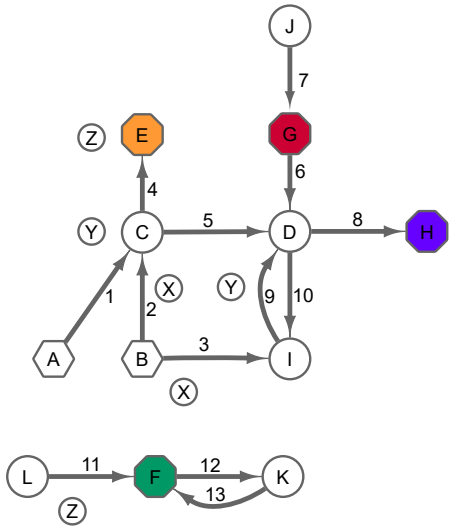
The targets represented by the red fractions are contained in components with start metabolites but do not have a necessary arc path from a start metabolite to the target, such as riboflavin (C00255).

The targets represented by the purple fractions are connected to a potential start metabolite in the network via an arc path. However, this is not sufficient for a valid pathway candidate. In addition, the arcs have to be associated with reactions for which all substrates are available or producible to ensure that the pathway candidate is feasible (BLASS et al., 2017). However, for these targets there is no reaction in the set of feasible reactions (see Section 5.2 Model Building) that produces that target for the last one arc of the arc path, which means that the overall pathway is not feasible. Note that the other arc-reaction associations thus do not matter in this case. An example for such a target is biotin (C00120) (Appendix B.4).

The targets represented by the brown fractions are targets that are predicted to have pathway candidates as they are connected to predefined start metabolites by an arc path and are produced by feasible reactions. However, our path-finding algorithm could not determine valid pathway candidates. To explore the reasons for this, we list the feasible reactions of the respective models that produce these targets for each of those targets. For each of these reactions we determine why it is not part of a pathway candidate. We identified three non-disjoint categories in which we can sort these reactions. To the first category belong reactions that produce the target but do not have arcs containing the target. As discussed in Section 5.2 Path-finding, a valid pathway candidate has to include a reaction with an arc to the target. Reactions that produce the targets only from substrates that are designated cofactors or inorganic metabolites are also sorted to this category, as they are correctly predicted to be feasible. However, our path-finding method does not handle such pathways since a valid pathway candidate requires at least one arc by definition and there are no arcs containing cofactors and inorganic metabolites. The second category encompasses reactions that do have an arc to the target, but require a substrate that is also a target for which no pathway candidate has been identified with our method. The reactions in the third category cannot be used in a pathway candidate because of a constraint in the MILP, which excludes pathways that use reactions consuming the target, as discussed in Section 5.2

Path-finding. There is no valid sequence of reactions with arcs that is feasible without using supplying reactions that consume the target. An example for a target of the brown fraction is 5'-methylthioadenosine (C00170), where the reactions producing this target belong to the first two categories discussed above (Appendix B.4).

To illustrate the different target categories, the example arc graph in Figure 5.7(a) depicts examples for each of the categories. Note that, for the sake of clarity, the depicted arc graph has additional vertices for the cofactors (small circles), which would normally not be part of the graph. The potential start metabolites A and B are depicted by hexagons, the potential targets E, F, G and H by octagons. Figure 5.7(b) lists the reaction equations and the arcs belonging to these reactions. A valid pathway candidate to E consists of the reactions R1 and



(a) Arc graph

reaction	equation	arcs
R1	$A + B \rightarrow C$	1,2
R2	$B + X \rightarrow I$	3
R3	$C + Y \rightarrow E + Z$	4
R4	$C + X \rightarrow D$	5
R5	$G \rightarrow D$	6
R6	$J \rightarrow G$	7
R7	$D + W \rightarrow H$	8
R8	$I + X \leftrightarrow D + Y$	9,10
R9	$L + Z \rightarrow F$	11
R10	$F \leftrightarrow K$	12,13

(b) Reaction definitions

Figure 5.7: Arc graph with examples for the different target categories in Figure 5.6 and the corresponding reactions. (a): Small circles: cofactor metabolites; hexagons: potential start metabolites; octagons: potential targets; large circles: metabolites not in any of the previous categories. Orange: target of the orange category; green: target of the green category; red: target of the red category; purple: target of the purple category. The numbers on the arcs refer to the column 'arcs' in (b). (b): Reactions in the example network with their respective reaction equations.

R3 (arcs 1, 2 and 4), since all needed substrates, i.e. A, B and Y, are available. E would thus be a target represented by the orange fractions in Figure 5.6. For target F it is not possible to find a pathway candidate because F is part of a graph component that does not include potential start metabolites. F is thus an example of the green fractions. Target G is part of the component which also includes the potential start metabolites. However, there is no arc

path connecting metabolites A or B to G, which makes G a target represented by the red fractions. Target H is an example for the purple fraction. To reach H, there are valid arc paths (e. g. $1 \rightarrow 5 \rightarrow 8$ or $2 \rightarrow 5 \rightarrow 8$), however the last reaction belonging to arc 8 requires W as a substrate, which is not available.

5.4 Concluding Remarks

Our presented method allows creating and characterizing genome-scale metabolic network reconstructions for the planning of biosynthetic production pathways using cell-free systems. The data are taken from biological databases, e.g. KEGG. We also discussed typical problems in the context of network reconstruction and how these can be solved in order to obtain applicable network models. We used the presented method for establishing models for the network reconstruction of a pan-organism from the whole KEGG database as well as for several interesting model organisms. We also used our path-finding method based on a global optimization problem to compute pathway candidates for all possible target molecules in the models and demonstrated that our method yields correct and meaningful results and that it is widely applicable for all kinds of networks and network sizes. The increasing availability of larger-scale metabolic networks that are increasingly well curated, as is e.g. already the case for *E. coli* and *S. cerevisiae* (ORTH et al., 2011; ZOMORRODI et al., 2010), will also increase the power of our method. Our network analysis method for multi-enzyme systems that do not have any cellular compartments particularly lacking a cell membrane differs significantly from published methods for whole cells with a defined link to the extracellular environment via transport systems (von KAMP et al., 2017; S.-Z. WANG et al., 2017) or with models of microbial communities, e.g. (MAGNÚSDÓTTIR et al., 2018).

The tools we presented are directly applicable to designing the synthesis of target compounds in cell-free systems. Our analysis tools - especially the feasibility prediction we described in Section 5.2 Model Building - are useful tools to predict if a target could potentially be produced in a given model and could thus be used to quickly screen if a host organism or strain is potentially capable of producing a certain product directly. If this is not the case, a comparison of biosynthesis pathways in a selected host organism and in the pan-organism is useful for identifying genetic engineering targets to create a production organism eventually. Our tools help identifying heterologous enzymes that might be candidates for insertion in the host organism chosen using genetic engineering to complete a desired pathway in that organism. Our tools also help to answer which substrates are required for a certain synthesis pathway.

Biosynthesis pathway candidates including stoichiometric and thermodynamic constraints can be determined with our presented path-finding algorithm presented earlier (BLASS et al., 2017). As reviewed in a recent publication (LIN et al., 2019), various methods have already been published and are in development that additionally allow the identification of new reactions considering the promiscuity of many enzymes but also the chemical similarity of substrates of these enzymes.

Our network reconstructions are the basis for the identification of gaps in the network that would prohibit synthesis of a desired target. With our tools, it is possible to identify potential gap fillers from the pan-organism network, which can then be implemented in an organism of interest using genetic engineering. It is also possible to do manual directed gap filling in the pan-organism network, e.g. by considering generic reactions, reactions not contained in KEGG, or expert reasoning.

Overall, our tools and networks are a suitable basis for focused and directed experimental work and the implementation of the synthesis of target compounds in cell-free systems.

CHAPTER 6

Synthesis Paths for UDP-glucose

In this chapter, the synthesis of UDP-glucose is discussed as an example for the usage of the presented path-finding tool for the design of synthesis pathways. UDP-glucose is a nucleotide sugar that plays an important role as an intermediate in various metabolic pathways (RALEVIC, 2015). The path-finding tool is used to computationally search for a synthesis pathway for the already experimentally shown multi-step synthesis of UDP-glucose from sucrose, UMP, ATP and phosphate (WEYLER et al., 2015) in recombinant *E. coli*, shown in Figure 6.1. The

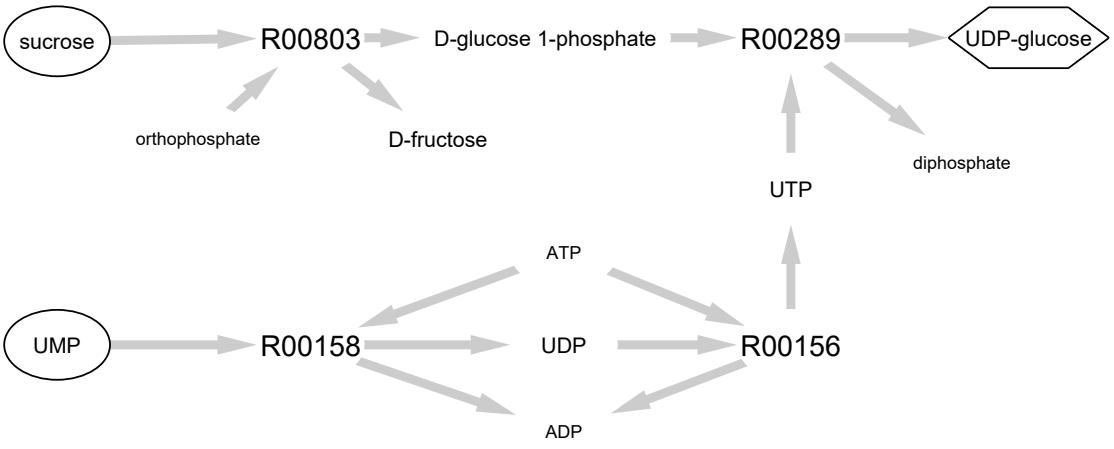


Figure 6.1: Pathway for the synthesis of UDP-glucose (WEYLER et al., 2015). R00803: sucrose:phosphate α-D-glucosyltransferase; R00289: UTP:α-D-glucose-1-phosphate uridylyltransferase; R00156: ATP:UDP phosphotransferase; R00158: ATP:UMP phosphotransferase. Metabolite names in small script denote cofactors or inorganic metabolites.

pathway consists of two branches. The first branch starting from sucrose via D-glucose-1-phosphate to UDP-glucose and the second branch starting from UMP via UDP and UTP to UDP-glucose.

6.1 Model

The pan-organism network model for the design of the synthesis pathway is the same model that has been used in the experiments discussed in Chapter 5 Network Reconstructions for Cell-Free Systems with two changes. The first change is the addition of UMP to the hand-curated list of basis metabolites presented in Section 5.2 Model Building. The second change is the removal of UTP from the list of cofactors and inorganic compounds to allow for the synthesis of UTP from UDP.

6.2 Path-Finding and Pathway Candidates

As the longest branch of the experimentally implemented pathway has length three, the path-finding tool is used to exhaustively find pathway candidates with at most three reactions on the linear path. The search results in 1203 pathway candidates. Upon inspection of the candidates, one can observe candidates which involve a direct cycle formed by supplying reactions that use metabolites that are not part of the metabolite pool and would thus not be possible in real world applications without the addition of those metabolites to the medium. Such pathway candidates have already been discussed in Section 5.3 Reachability Analysis (blue fractions of Figure 5.6). As a global optimization method, the implemented MILP does not detect such cycles. However, a filter step to remove such pathway candidates has been developed. The filter takes the set of active reactions of the pathway and the metabolite pool. It tests for each reaction if all substrates of the reaction are contained in this pool. If this is the case, the products of the respective reaction are added to the metabolite pool. The active reactions are tested until no new metabolites are added. If there remain any active reactions that are not feasible with the substrates of the metabolite pool, the whole pathway is not feasible and thus filtered from the set of pathway candidates.

After filtering 442 pathway candidates remain. The shortest pathway candidate consists of two active reactions, while the longest has 39 active reactions. The pathway candidates have 117 unique linear paths with different supplying reactions. The start metabolites of the linear pathways are listed in Table 6.1 together with the number of pathway candidates from the respective start metabolite. The table shows that most pathway candidates start

with D-glucose (12 candidates), followed by UMP (10 candidates). For sucrose, there are five pathway candidates.

Table 6.1: Start metabolites of the pathway candidates to UDP-glucose with KEGG ids and the number of pathway candidates.

start metabolite	KEGG id	number of pathway candidates
D-glucose	C00031	12
UMP	C00105	10
D-glucose 6-phosphate	C00092	7
D-fructose	C00095	5
pyruvate	C00022	5
sucrose	C00089	5
α-D-glucose	C00267	4
2-oxoglutarate	C00026	4
acetate	C00033	4
citrate	C00158	4
oxaloacetate	C00036	4
L-serine	C00065	4
β-D-glucose	C00221	3
glycerol	C00116	3
glycine	C00037	3
malate	C00149	3
L-alanine	C00041	3
L-arginine	C00062	3
L-aspartate	C00049	3
L-cysteine	C00097	3
L-glutamate	C00025	3
L-glutamine	C00064	3
L-histidine	C00135	3
D-ribose	C00121	2
fumarate	C00122	1
L-lysine	C00047	2
L-threonine	C00188	2
succinate	C00042	2
L-phenylalanine	C00079	1
L-proline	C00148	1
L-tryptophan	C00078	1
L-tyrosine	C00082	1
L-valine	C00183	1

The shortest pathway proposed by the path-finding tool that can produce UDP-glucose from the given metabolite pool is the two-step pathway shown in Figure 6.2. This pathway

produces UDP-glucose from sucrose, UMP and ATP. The linear path of the pathway is sucrose → UDP-glucose (black arrows). UDP is synthesized from UMP and ATP by reaction R00158. Listing 1 in Appendix Section C.1 shows the details of this pathway candidate.

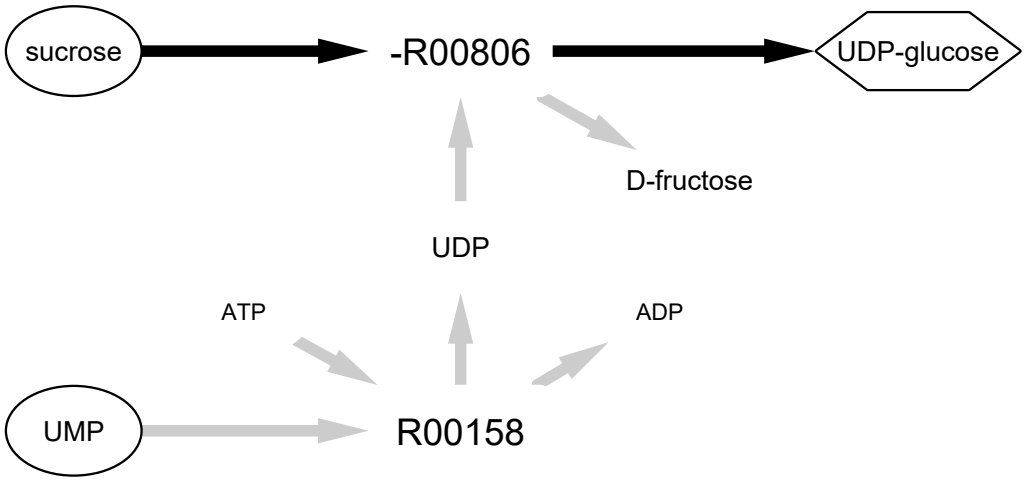


Figure 6.2: Two-step pathway candidate to UDP-glucose from sucrose, UMP and ATP. R00806: UDP-glucose:D-fructose 2-α-D-glucosyltransferase; R00158: ATP:UMP phosphotransferase. Black arrows: arcs used on the linear path. Metabolite names in small script denote cofactors or inorganic metabolites.

Figure 6.3 depicts the second shortest pathway candidate, consisting of three active reactions. It synthesizes UDP-glucose from sucrose, Cytidine 5'-triphosphate (CTP), orthophosphate and water. The linear path of this pathway is sucrose → D-glucose 1-phosphate → UDP-glucose, designated by black arrows. It corresponds to the branch starting with sucrose of the experimentally shown pathway (Figure 6.1). Reaction R00568 is the supplying reaction for UTP from CTP and water. The details of this pathway candidate are shown in Listing 2 of Appendix Section C.1.

The search also proposes a pathway candidate starting from UMP, D-glucose 6-phosphate and ATP, depicted in Figure 6.4 (details in Listing 3 of Appendix Section C.1). Its linear path is UMP → UDP → UTP → UDP-glucose (black arrows), corresponding to the UMP branch of the experimentally shown pathway (Figure 6.1).

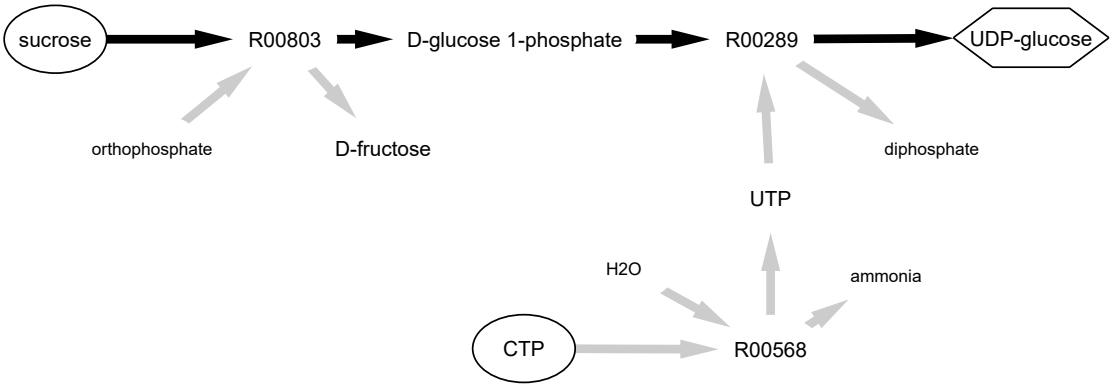


Figure 6.3: Pathway candidate to UDP-glucose from sucrose, CTP, orthophosphate and water. R00803: sucrose:phosphate α-D-glucose-glucosyltransferase; R00289: UTP:α-D-glucose-glucose-1-phosphate uridylyltransferase; R00568: CTP aminohydrolase. Black arrows: arcs used on the linear path. Metabolite names in small script denote cofactors or inorganic metabolites.

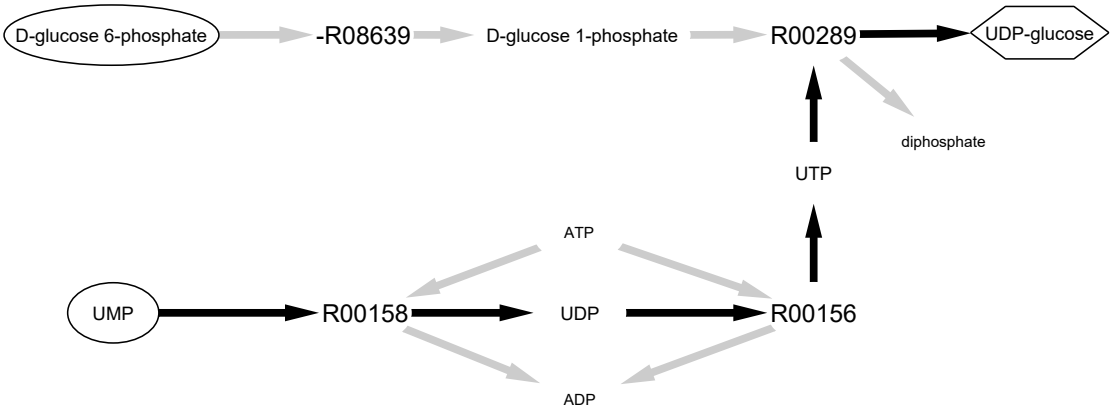


Figure 6.4: Pathway candidate to UDP-glucose from D-glucose 6-phosphate, UMP and ATP. R08639: α-D-glucose 1,6-phosphomutase; R00289: UTP:α-D-glucose-glucose-1-phosphate uridylyltransferase; R00156: ATP:UDP phosphotransferase; R00158: ATP:UMP phosphotransferase. Black arrows: arcs used on the linear path. Metabolite names in small script denote cofactors or inorganic metabolites.

6.3 Discussion

As the presented results show, the experimentally implemented pathway depicted in Figure 6.1 can be obtained by combining the linear paths of the pathway candidates shown in Figures 6.3 and 6.4. The combination of multiple pathways is an interesting approach to generate further pathway candidates for cases in which the desired pathway is not contained in the path-finding results. It can be applied whenever reactions with multiple arcs such as reaction R00289 in Figure 6.1 are included in a pathway. This reaction contains the arcs D-glucose-1-phosphate - UDP-glucose and UTP - UDP-glucose and therefore allows the combination of pathways were one of those arcs is active with pathways where the other arc is active.

To explain why the tool did not propose the pathway from Figure 6.1 as a single pathway candidate, the details of the path-finding algorithm have to be inspected in more detail. As already described in Section 4.2 MILP, a pathway consists of the linear path and supplying reactions. In the pathway candidate shown in Figure 6.4, the linear path is UMP → UDP → UTP → UDP-glucose (R00158 → R00156 → R00289). D-glucose 1-phosphate, which is a required substrate for R00289, is not contained in the metabolite pool. It thus has to be synthesized. In this pathway candidate, this is done by reaction -R08639 (the supplying reaction) from D-glucose 6-phosphate. Reaction R00803, which is used in the experimentally implemented pathway, is not found as an alternative since the combination of active arcs of the path and the chosen reactions are not different and a full enumeration of the supplying reactions in the MILP has not been implemented. Since for the optimization problem all reactions are treated equally, the choice of supplying reactions is arbitrarily done by the CPLEX solver. A means of influencing which supplying reactions the solver chooses for a pathway candidate besides minimizing the number of supplying reactions has not been included in the algorithm. A possible extension of the objective function of the MILP that heuristically takes into account the $\Delta_r G$ value of a reaction by preferring reactions with a smaller $\Delta_r G$ can be implemented as shown in Equation (6.1).

$$Minimize \quad \sum_{i=1}^{|M|} \sum_{j=1, j\neq i}^{|M|} u_{ij} + \frac{1}{2 \cdot |R| + 1} \sum_{i=1}^{|R|} (1 + \Delta Gn_i) \cdot z_i \tag{6.1}$$

ΔGn_i is the $\Delta_r G$ of reaction i normalized by the maximum $|\Delta_r G|$ of all reactions in the network reconstruction. This objective function minimizes the number of active reactions and the $\Delta_r G$ of the supplying reactions. Using this objective function in the path-finding, the pathway shown in Figure 6.1 can be directly retrieved as the second pathway candidate.

Part IV

Conclusion

CHAPTER 7

Extended Summary

This study presents the findings of the work on the two main aims in the scope of cell-free systems defined in Chapter 3 Aims and Scope. A comprehensive workflow for the directed design of biosynthetic production pathways using cell-free systems and the planning of such syntheses (Part II Path-Finding and Network Analysis for Multi-Enzyme Biocatalysis) has been developed and discussed. Additionally, a method for the creation and characterization of genome-scale metabolic network reconstructions for the planning of biosynthetic production pathways (Part III Network Reconstructions for Cell-Free Systems) has been established.

7.1 Path-Finding and Ranking

The developed path-finding method presented in this work is suitable for finding synthesis pathways in metabolic network reconstructions of cell-free systems. The MILP-based method computes pathway candidates to a given target metabolite in a metabolite graph, in which the metabolites are connected by arcs derived from biologically meaningful reaction pairs defined in the KEGG RCLASS database. A pathway discovered by the method consists of two parts. The first part is a sequence of metabolites connected by reactions starting with one of the start metabolites and ending with the product. The second part is a minimal set of reactions supplying substrates required by the pathway reactions that are not directly available in the metabolite pool composed of start metabolites, cofactors and inorganic compounds. The MILP is defined by constraints that take into account the network topology and the stoichiometries of the underlying reactions to identify biologically meaningful pathway candidates. The constraints of the MILP can be adapted freely to customize it for the task at hand. The presented method stands in contrast to other methods that either only account

for shortest paths in a graph between a given start and end node without taking into account stoichiometry; or methods that depend on a steady state in the network, which is not the case in cell-free systems.

To handle the large number of results in the broad solution space of a pathway search to a target metabolite and to highlight the most meaningful candidates, the pathways generated are ranked according to various different criteria. The criteria comprise metrics such as pathway length, reaction thermodynamics, the number of heterologous enzymes in a given host organism, cofactor requirement, or number of potential side reactions. However, these criteria can be fully adapted and expanded, to take into account further aspects that might be of importance for a given synthesis.

By means of the examples GPP, amygdalin, pyrrolysine and (S)-2-phenyloxirane it is shown that the method proposes meaningful pathway candidates that can be used as a base for further investigation to find the most promising synthesis pathway candidate. The synthesis pathway for the multi-step synthesis of UDP-glucose from sucrose, UMP, ATP and phosphate that was implemented in a recombinant, permeabilized *E. coli* strain (WEYLER et al., 2015) (Chapter 6 Synthesis Paths for UDP-glucose) has also been recovered successfully.

Overall, the method presented in this work is a useful tool for the planning of biosynthetic syntheses. The different steps of the path-finding and ranking workflow can be adapted to meet the needs of a specific project at hand, which makes the method highly versatile and suitable for a variety of problems.

7.2 Model Building and Analysis

A workflow for building and characterizing genome-scale metabolic network reconstructions for the planning of biosynthetic production pathways has been developed. This workflow is used for the reconstruction of networks from the KEGG databases COMPOUND, REACTION and ENZYME. The large number of reactions in KEGG has to be filtered in order to include only meaningful reactions into our model reconstructions. For the network models, only reactions whose reactants have KEGG COMPOUND identifiers, integer stoichiometric coefficients, that are not generic or contain generic reactants and that have reaction class annotations are taken into account. Reactions satisfying these requirements are included in a pan-organism network model. Additionally, models from nine specific organisms chosen for biotechnological and scientific importance are compiled using the organism annotation for the genes of the enzymes catalyzing the suitable reactions.

The metabolites in the models are grouped into different categories, which are treated differently in the path-finding algorithm. Potential start metabolites can be used as start point for a pathway candidate. This is also true for basis metabolites, which constitute a hand-curated subset of inexpensive and easily available metabolites. Cofactors and inorganic metabolites are excluded from arcs, but are freely available as substrates for reactions in a pathway. These sets constitute a metabolite pool of freely available metabolites.

For all models, all possible target metabolites are determined automatically and pathway candidates for them are computed. The presented path-finding method yields meaningful results in different kinds of networks and network sizes. A tool based on BFS to quickly predict if a given target compound could potentially be produced in a given host organism has also been implemented. The different properties of the network reconstructions, such as network hubs and connected components are furthermore analyzed. These properties, together with the target reachability analysis allow for a more in-depth analysis of the networks.

CHAPTER 8

Concluding Remarks and Outlook

In the following, the aspects that have not been already addressed fully in other parts of this work are wrapped up and discussed. Furthermore, this chapter gives an outlook for subsequent expansions of the presented algorithm and methods.

8.1 Network Reconstruction and Curation

The data used in this study for assembling network reconstructions following the workflow presented in Chapter 2 Network Design is taken from KEGG. However, depending on the aim of the study in which the network reconstruction is used, it can be fully adapted to the needs by using any database or data source providing the necessary information (Section 2.1 Databases). In general, the quality of a network reconstruction depends heavily on the quality of the data used. It is thus important to use high quality (curated) data sources that contain as much information as needed to obtain network reconstructions that are as comprehensive as possible.

Gap filling of the network reconstruction is a vital step for obtaining well-founded results with respect to finding pathway candidates to a product. So in a further iteration of the network reconstruction workflow, it would be beneficial to employ gap-filling strategies to obtain even more comprehensive network reconstructions (Section 2.3 Network Reconstruction).

Almost 50% of the reaction entries in KEGG (version 65.0) are orphan reactions lacking an associated protein sequence or enzyme (SOROKINA et al., 2014). While building the organism-specific network model reconstructions it can be observed that only about 78 % of the reactions from the pan-organism model have associated enzyme(s) (Figure 5.3). As in

the presented filtering scheme enzyme associations are a requirement for associating the respective reaction to an organism, a large part of the reactions in KEGG does not qualify for an organism-specific model as these reactions do not have any enzymes associated. Associating orphan reactions with enzymes could thus improve the quality of network reconstruction models.

As discussed in Section 5.2 Model Building, the default parameters of eQuilibrator's (FLAMHOLZ et al., 2012) have been used for the computation of the $\Delta_r G$ values. To determine reaction reversibility, a $\Delta_r G$ value of 15 kJ/mol was chosen empirically. However, the parameters for the thermodynamics can be easily adapted if necessary, even individually for each reaction. Additionally, the estimated values could easily be replaced by experimental values, if available.

A meaningful expansion of the network reconstruction step would be to take into account enzyme concentrations or kinetic parameters (SRINIVASAN et al., 2015).

8.2 Path-Finding

The presented path-finding method is highly customizable and can easily be adapted to the respective requirements of the studies in which it is employed. There are numerous aspects of the tool that can be tailored to the respective needs. One possibility is to customize the MILP by adding, modifying or removing constraints. In this study, the MILP is applied on a network representing cell free systems instead of a living organism, so it is not assumed that any metabolite is in steady-state in the network. However, adding the constraint (8.1) (the mathematical notation following Section 4.2 MILP), would be applicable in the case where the objective is to find synthesis paths in living cells. Given a set of internal metabolites I, the constraint ensures that these metabolites are balanced.

$$\sum_{r=1}^{|R|} S_{mr} v_r = 0, \quad \forall m \in I \tag{8.1}$$

The result of the path-finding heavily depends on the choice of arcs. The reason for choosing reaction pairs as basis for the arcs in this study has been discussed in Appendix Section B.2. However, the method of extracting arcs from reactions is totally customizable and can be fully adapted.

It is also possible to completely customize the different metabolite lists. The start and basis metabolites can be specified depending on the specified target or on the availability of certain substrates. The list of cofactors and inorganic compounds is also fully customizable to account for a specific kind of medium composition.

The objective function of the MILP presented in Section 6 that minimizes the number of active reactions as well as the $\Delta_r G$ of the supplying reactions is only one possible extension of the path-finding algorithm. A further aspect that could be taken into account are the costs of substrates that are needed for the supplying reactions.

A further extension of the MILP would be to investigate if a full enumeration of the supplying reactions is possible (and meaningful) to deliver even more pathway candidates.

8.3 Ranking

All ranking criteria can be fully adapted and extended. It is possible to define all kinds of ranking criteria based on the respective synthesis. The order of the ranking criteria can be rearranged to change the impact of a specific criterion. Additionally, more ranking criteria can be incorporated. A practical and meaningful extension would be to incorporate the prices of the substrates. This was not possible during the study due to the lack of a readily available method for automatically extracting prices for purchasable substrates. Given such a price list, it would be possible to rank the pathway candidates based on the prices of their required substrates. One could even tailor the list of starting metabolites depending on the availability of certain substrates and exclude those that are not purchasable. A step forward to this ranking criterion would be to categorize the substrates into categories such as inexpensive/expensive and prefer pathway candidates that contain more inexpensive substrates.

8.4 Further Aspects

In Chapter 4 Network Design and Analysis for Multi-Enzyme Biocatalysis the detection of potential side reaction for a given pathway candidate as an additional ranking criterion is presented. However, this criterion can only consider reactions already incorporated in KEGG. In most cases, KEGG only contains the main reaction(s) for a specific enzyme. It would thus be a meaningful addition to take into account further data sources and tools which contain information on side reactions of enzymes, such as BRENDA (JESKE et al., 2018), MINE (JEFFRYES et al., 2015) or ATLAS of Biochemistry (HADADI et al., 2016).

A further interesting extension would be to account for cofactor regeneration systems. This could for example be done by taking into account cofactor usage and adding reactions that regenerate cofactors to a pathway candidate.

Bibliography

AEHLE, W. (2004): *Enzymes in industry*. Wiley-VCH. (Cit. on p. 4).

AI, C., & KONG, L. (2018): CGPS: A machine learning-based approach integrating multiple gene set analysis tools for better prioritization of biologically relevant pathways. *Journal of Genetics and Genomics* (9 2018), vol. *45*: 489–504. https://doi.org/10.1016/j.jgg.2018.08.002 (cit. on p. 18)

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W., & LIPMAN, D. J. (1990): Basic local alignment search tool. *Journal of molecular biology* (1990), vol. *215*([3]): 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2 (cit. on p. 22)

ARTIMO, P., JONNALAGEDDA, M., ARNOLD, K., BARATIN, D., CSARDI, G., de CASTRO, E., DUVAUD, S., FLEGEL, V., FORTIER, A., GASTEIGER, E., GROSDIDIER, A., HERNANDEZ, C., IOANNIDIS, V., KUZNETSOV, D., LIECHTI, R., MORETTI, S., MOSTAGUIR, K., REDASCHI, N., ROSSIER, G., ... STOCKINGER, H. (2012): ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Research* (2012), vol. *40*([W1]): W597–W603. https://doi.org/10.1093/Nar/Gks400 (cit. on p. 16)

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T., HARRIS, M. A., HILL, D. P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J. C., RICHARDSON, J. E., RINGWALD, M., RUBIN, G. M., ... GENE ONTOLOGY, C. (2000): Gene ontology: Tool for the unification of biology. *Nature Genetics* (2000), vol. *25*([1]): 25–29 (cit. on p. 12).

BAIROCH, A. (2000): The enzyme database in 2000. *Nucleic Acids Research* (2000), vol. *28*([1]): 304–305. https://doi.org/10.1093/nar/28.1.304 (cit. on p. 13)

BATES, J. T., CHIVIAN, D., & ARKIN, A. P. (2011): GLAMM: Genome-linked application for metabolic maps. *Nucleic Acids Research* (2011), vol. *39*: W400–W405. https://doi.org/10.1093/nar/gkr433 (cit. on pp. 18, 20)

BECKER, J., GIESSELMANN, G., HOFFMANN, S. L., & WITTMANN, C. (2016): *Corynebacterium glutamicum* for sustainable bioproduction: From metabolic physiology to systems metabolic engineering. *Synthetic biology–metabolic engineering* (pp. 217–263). Springer. (Cit. on p. 70).

BENSON, D. A., CAVANAUGH, M., CLARK, K., KARSCH-MIZRACHI, I., LIPMAN, D. J., OSTELL, J., & SAYERS, E. W. (2013): Genbank. *Nucleic acids research* (2013), vol. *41*([D1]): D36–42. https://doi.org/10.1093/nar/gks1195 (cit. on p. 12)

BLASS, L. K., WEYLER, C., & HEINZLE, E. (2017): Network design and analysis for multi-enzyme biocatalysis. *BMC Bioinformatics* (Aug. 2017), vol. *18*([1]): 366. https://doi.org/10.1186/s12859-017-1773-y (cit. on pp. 40, 61, 62, 77, 80, 123)

BLAZECK, J., & ALPER, H. (2010): Systems metabolic engineering: Genome-scale models and beyond. *Biotechnology Journal* (2010), vol. *5*([7]). https://doi.org/10.1002/biot.200900247 (cit. on p. 28)

BLUM, T., & KOHLBACHER, O. (2008a): MetaRoute: fast search for relevant metabolic routes for interactive network navigation and visualization. *Bioinformatics* (2008), vol. *24*([18]): 2108–2109 (cit. on p. 37).

BLUM, T., & KOHLBACHER, O. (2008b): Using atom mapping rules for an improved detection of relevant routes in weighted metabolic networks. *Journal of Computational Biology* (2008), vol. *15*([6]): 565–576 (cit. on p. 37).

BOGMAN, K., ERNE-BRAND, F., ALSENZ, J., & DREWE, J. (2003): The role of surfactants in the reversal of active transport mediated by multidrug resistance proteins. *Journal of Pharmaceutical Sciences* (2003), vol. *92*([6]): 1250–1261. https://doi.org/10.1002/jps.10395 (cit. on p. 30)

BORNSCHEUER, U. T., HUISMAN, G. W., KAZLAUSKAS, R. J., LUTZ, S., MOORE, J. C., & ROBINS, K. (2012): Engineering the third wave of biocatalysis. *Nature* (2012), vol. *485*([7397]): 185–194. https://doi.org/Doi10.1038/Nature11117 (cit. on pp. 3, 4)

BOWERS, P. M., PELLEGRINI, M., THOMPSON, M. J., FIERRO, J., YEATES, T. O., & EISENBERG, D. (2004): Prolinks: A database of protein functional linkages derived from coevolution. *Genome Biology* (2004), vol. *5*([5]). https://doi.org/10.1186/gb-2004-5-5-r35 (cit. on p. 13)

BUJARA, M., & PANKE, S. (2012): In silico assessment of cell-free systems. *Biotechnology and Bioengineering* (2012), vol. *109*([10]): 2620–2629. https://doi.org/Doi10.1002/Bit.24534 (cit. on p. 9)

CARBONELL, P., PLANSON, A.-G., FICHERA, D., & FAULON, J.-L. (2011): A retrosynthetic biology approach to metabolic pathway design for therapeutic production. *BMC Systems Biology* (2011), vol. *5*([1]): 122 (cit. on p. 37).

CARSTEN, J. M., SCHMIDT, A., & SIEBER, V. (2015): Characterization of recombinantly expressed dihydroxy-acid dehydratase from *Sulfobus solfataricus* -a key enzyme for the conversion of carbohydrates into chemicals. *Journal of biotechnology* (2015), vol. *211*: 31–41. https://doi.org/10.1016/j.jbiotec.2015.06.384 (cit. on p. 37)

CHELLIAH, V., JUTY, N., AJMERA, I., ALI, R., DUMOUSSEAU, M., GLONT, M., HUCKA, M., JALOWICKI, G., KEATING, S., KNIGHT-SCHRIJVER, V., Et al. (2014): BioModels: Ten-year anniversary. *Nucleic acids research* (2014), vol. *43*([D1]): D542–D548 (cit. on pp. 14, 15).

CHERRY, J. M., HONG, E. L., AMUNDSEN, C., BALAKRISHNAN, R., BINKLEY, G., CHAN, E. T., CHRISTIE, K. R., COSTANZO, M. C., DWIGHT, S. S., ENGEL, S. R., Et al. (2011): Saccharomyces genome database: The genomics resource of budding yeast. *Nucleic acids research* (2011), vol. *40*([D1]): D700–D705 (cit. on pp. 14, 17).

CONSORTIUM, T. G. O. (2019): The gene ontology resource: 20 years and still going strong. *Nucleic acids research* (D1 Jan. 2019), vol. *47*: D330–D338. https://doi.org/10.1093/nar/gky1055 (cit. on p. 12)

CONSORTIUM, U. (2018): Uniprot: A worldwide hub of protein knowledge. *Nucleic acids research* (2018), vol. *47*([D1]): D506–D515 (cit. on p. 13).

CORMEN, T. H., LEISERSON, C. E., RIVEST, R. L., & STEIN, C. (2009): *Introduction to algorithms*. MIT press. (Cit. on pp. 29, 67, 72).

DEMIR, E., CARY, M. P., PALEY, S., FUKUDA, K., LEMER, C., VASTRIK, I., WU, G., D'EUSTACHIO, P., SCHAEFER, C., LUCIANO, J., Et al. (2010): The BioPAX Community Standard for Pathway Data Sharing. *Nature Biotechnology* (2010), vol. *28*([9]): 935–942. https://doi.org/10.1038/nbt.1666 (cit. on pp. 17, 28)

DEVILLE, Y., GILBERT, D., van HELDEN, J., & WODAK, S. J. (2003): An overview of data models for the analysis of biochemical pathways. *Briefings in bioinformatics* (3 Sept. 2003), vol. *4*: 246–259. https://doi.org/10.1093/bib/4.3.246 (cit. on p. 26)

DRAUZ, K., GRÖGER, H., & MAY, O. (2012): *Enzyme catalysis in organic synthesis, 3 volume set* (Vol. 1). John Wiley & Sons. (Cit. on p. 4).

DUDLEY, Q. M., ANDERSON, K. C., & JEWETT, M. C. (2016): Cell-free mixing of *Escherichia coli* crude extracts to prototype and rationally engineer high-titer mevalonate synthesis. *ACS Synthetic Biology* (2016), vol. *5*([12]): 1578–1588 (cit. on p. 37).

DUROT, M., BOURGUIGNON, P.-Y., & SCHACHTER, V. (2009): Genome-scale models of bacterial metabolism: Reconstruction and applications. *Fems Microbiology Reviews* (2009), vol. *33*([1]): 164–190. https://doi.org/10.1111/j.1574-6976.2008.00146.x (cit. on p. 29)

ELBOURNE, L. D., TETU, S. G., HASSAN, K. A., & PAULSEN, I. T. (2016): TransportDB 2.0: A database for exploring membrane transporters in sequenced genomes from all domains of life. *Nucleic acids research* (2016), vol. *45*([D1]): D320–D324. https://doi.org/10.1093/nar/gkw1068 (cit. on p. 11)

ENDO, T., & KOIZUMI, S. (2001): Microbial conversion with cofactor regeneration using genetically engineered bacteria. *Advanced Synthesis & Catalysis* (2001), vol. *343*([6-7]): 521–526. https://doi.org/10.1002/1615-4169(200108)343:6/7<521::AID-ADSC521>3.0.CO;2-5 (cit. on p. 60)

FABREGAT, A., JUPE, S., MATTHEWS, L., SIDIROPOULOS, K., GILLESPIE, M., GARAPATI, P., HAW, R., JASSAL, B., KORNINGER, F., MAY, B., Et al. (2017): The reactome pathway knowledgebase. *Nucleic acids research* (2017), vol. *46*([D1]): D649–D655 (cit. on pp. 16, 17).

FAILMEZGER, J., SCHOLZ, S., BLOMBACH, B., & SIEMANN-HERZBERG, M. (2018): Cell-free protein synthesis from fast-growing *Vibrio natriegens*. *Frontiers in microbiology* (2018), vol. *9*: 1146. https://doi.org/10.3389/fmicb.2018.01146 (cit. on p. 70)

FAUST, K., CROES, D., & van HELDEN, J. (2009): Metabolic pathfinding using RPAIR annotation. *Journal of molecular biology* (2009), vol. *388*([2]): 390–414 (cit. on p. 37).

FEIST, A. M., HENRY, C. S., REED, J. L., KRUMMENACKER, M., JOYCE, A. R., KARP, P. D., BROADBELT, L. J., HATZIMANIKATIS, V., & PALSSON, B. O. (2007): A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Molecular Systems Biology* (2007), vol. *3*: 18. https://doi.org/10.1038/msb4100155 (cit. on p. 9)

FELIX, H. (1982): Permeabilized cells. *Analytical Biochemistry* (1982), vol. *120*([2]): 211–234. https://doi.org/10.1016/0003-2697(82)90340-2 (cit. on p. 6)

FLAMHOLZ, A., NOOR, E., BAR-EVEN, A., & MILO, R. (2012): eQuilibrator - the biochemical thermodynamics calculator. *Nucleic Acids Research* (2012), vol. *40*([D1]): D770–D775 (cit. on pp. 22, 39, 64, 67, 94, 183).

GANTER, M., BERNARD, T., MORETTI, S., STELLING, J., & PAGNI, M. (2013): MetaNetX.org: A website and repository for accessing, analysing and manipulating metabolic networks. *Bioinformatics* (2013), vol. *29*([6]): 815–816. https://doi.org/D0I10.1093/bioinformatics/btt036 (cit. on p. 17)

GAO, J., ELLIS, L. B. M., & WACKETT, L. P. (2010): The University of Minnesota Biocatalysis/Biodegradation Database: Improving public access. *Nucleic Acids Research* (2010), vol. *38*: D488–D491. https://doi.org/10.1093/nar/gkp771 (cit. on p. 16)

GERARD, M. F., STEGMAYER, G., & MILONE, D. H. (2015): EvoMS: An evolutionary tool to find de novo metabolic pathways. *Biosystems* (2015), vol. *134*: 43–47. https://doi.org/10.1016/j.biosystems.2015.04.006 (cit. on p. 37)

GLONT, M., NGUYEN, T. V. N., GRAESSLIN, M., HÄLKE, R., ALI, R., SCHRAMM, J., WIMALARATNE, S. M., KOTHAMACHU, V. B., RODRIGUEZ, N., SWAT, M. J., Et al. (2017): BioModels: Expanding horizons to include more modelling approaches and formats. *Nucleic acids research* (2017), vol. *46*([D1]): D1248–D1253 (cit. on pp. 14, 15).

GROEGER, H., & HUMMEL, W. (2014): Combining the 'two worlds' of chemocatalysis and biocatalysis towards multi-step one-pot processes in aqueous media. *Current opinion in chemical biology* (2014), vol. *19*: 171–179. https://doi.org/10.1016/j.cbpa.2014.03.002 (cit. on p. 60)

HADADI, N., HAFNER, J., SHAJKOFCI, A., ZISAKI, A., & HATZIMANIKATIS, V. (2016): ATLAS of biochemistry: A repository of all possible biochemical reactions for synthetic biology and metabolic engineering studies. *ACS synthetic biology* (2016), vol. *5*([10]): 1155–1166 (cit. on pp. 12, 24, 95).

HADADI, N., MOHAMMADIPEYHANI, H., MISKOVIC, L., SEIJO, M., & HATZIMANIKATIS, V. (2019): Enzyme annotation for orphan and novel reactions using knowledge of substrate reactive sites. *Proceedings of the National Academy of Sciences* (2019), vol. *116*([15]): 7298–7307 (cit. on p. 24).

HAGGART, C. R., BARTELL, J. A., SAUCERMAN, J. J., & PAPIN, J. A. (2011): Whole-genome metabolic network reconstruction and constraint-based modeling. D. JAMESON, M. VERMA, & H. V. WESTERHOFF (Eds.), *Methods in enzymology, vol 500: Methods in systems biology* (pp. 411–433). San Diego, Elsevier Academic Press Inc. https://doi.org/10.1016/b978-0-12-385118-5.00021-9. (Cit. on pp. 22, 24, 29, 30)

HASTINGS, J., OWEN, G., DEKKER, A., ENNIS, M., KALE, N., MUTHUKRISHNAN, V., TURNER, S., SWAINSTON, N., MENDES, P., & STEINBECK, C. (2015): ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic acids research* (2015), vol. *44*([D1]): D1214–D1219 (cit. on pp. 10, 11).

HATZIMANIKATIS, V., LI, C., IONITA, J. A., HENRY, C. S., JANKOWSKI, M. D., & BROADBELT, L. J. (2005): Exploring the diversity of complex metabolic networks. *Bioinformatics* (2005), vol. *21*([8]): 1603–1609 (cit. on pp. 12, 37).

HEINZLE, E., BIWER, A. P., & COONEY, C. L. (2006): *Development of sustainable bioprocesses*. Wiley-VCH. (Cit. on p. 4).

HEINZLE, E., WEYLER, C., KRAUSER, S., & BLASS, L. K. (2013): Directed multistep biocatalysis using tailored permeabilized cells. A.-P. ZENG (Ed.), *Advances in biochemi-*

cal engineering/biotechnology (pp. 185–234). Springer Berlin Heidelberg. https://doi.org/10.1007/10_2013_240. (Cit. on pp. 36, 60)

HEIRENDT, L., ARRECKX, S., PFAU, T., MENDOZA, S. N., RICHELLE, A., HEINKEN, A., HARALDSDOTTIR, H. S., WACHOWIAK, J., KEATING, S. M., VLASOV, V., Et al. (2019): Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v. 3.0. *Nature protocols* (2019), vol.: 1 (cit. on p. 18).

HOFFART, E., GRENZ, S., LANGE, J., NITSCHEL, R., MÜLLER, F., SCHWENTNER, A., FEITH, A., LENFERS-LÜCKER, M., TAKORS, R., & BLOMBACH, B. (2017): High substrate uptake rates empower *Vibrio natriegens* as production host for industrial biotechnology. *Appl. Environ. Microbiol.* (2017), vol. *83*([22]): e01614–17. https://doi.org/10.1128/AEM.01614-17 (cit. on p. 70)

HORINOUCHI, N., OGAWA, J., KAWANO, T., SAKAI, T., SAITO, K., MATSUMOTO, S., SASAKI, M., MIKAMI, Y., & SHIMIZU, S. (2006): Efficient production of 2-deoxyribose 5-phosphate from glucose and acetaldehyde by coupling of the alcoholic fermentation system of baker's yeast and deoxyriboaldolase-expressing *Escherichia coli*. *Bioscience Biotechnology and Biochemistry* (2006), vol. *70*([6]): 1371–1378. https://doi.org/10.1271/bbb.50648 (cit. on p. 28)

HORINOUCHI, N., SAKAI, T., KAWANO, T., MATSUMOTO, S., SASAKI, M., HIBI, M., SHIMA, J., SHIMIZU, S., & OGAWA, J. (2012): Construction of microbial platform for an energy-requiring bioprocess: Practical 2'-deoxyribonucleoside production involving a C-C coupling reaction with high energy substrates. *Microbial Cell Factories* (2012), vol. *11*. https://doi.org/10.1186/1475-2859-11-82 (cit. on p. 28)

HUANG, Y., ZHONG, C., LIN, H. X., & WANG, J. (2017): A Method for Finding Metabolic Pathways Using Atomic Group Tracking. *PloS one* (2017), vol. *12*([1]): e0168725 (cit. on p. 37).

HUCKA, M., FINNEY, A., SAURO, H. M., BOLOURI, H., DOYLE, J. C., KITANO, H., ARKIN, A. P., BORNSTEIN, B. J., BRAY, D., CORNISH-BOWDEN, A., Et al. (2003): The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models. *Bioinformatics* (2003), vol. *19*([4]): 524–531 (cit. on p. 48).

HUNTER, J. D. (2007): Matplotlib: A 2D graphics environment. *Computing In Science & Engineering* (2007), vol. *9*([3]): 90–95. https://doi.org/10.1109/MCSE.2007.55 (cit. on p. 67)

HUSSEIN, A., CERDEÑO-TÁRRAGA, A., TORIBIO, A.-L., MILANO, A., ALAKO, B., AMID, C., SMIRNOV, D., RICHARDS, E., COCHRANE, G., CLELAND, I., RAJAN, J., MARTÍNEZ-VILLACORTA, J., REDDY, K., ROSELLO, M., SILVESTER, N., PAKSERESHT, N., LEINONEN, R., HOLT, S., VIJAYARAJA, S., ... HARRISON, P. W. (2018): The European Nucleotide

Archive in 2018. *Nucleic Acids Research* (Nov. 2018), vol. *47*([D1]): D84–D88. https://doi.org/10.1093/nar/gky1078 (cit. on p. 12)

JASSAL, B., MATTHEWS, L., VITERI, G., GONG, C., LORENTE, P., FABREGAT, A., SIDIROPOULOS, K., COOK, J., GILLESPIE, M., HAW, R., Et al. (2019): The reactome pathway knowledgebase. *Nucleic acids research* (2019), vol. *48*([D1]): D498–D503 (cit. on pp. 16, 17).

JEFFRYES, J. G., COLASTANI, R. L., ELBADAWI-SIDHU, M., KIND, T., NIEHAUS, T. D., BROADBELT, L. J., HANSON, A. D., FIEHN, O., TYO, K. E., & HENRY, C. S. (2015): MINEs: Open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. *Journal of cheminformatics* (2015), vol. *7*([1]): 44 (cit. on pp. 11, 95).

JESKE, L., PLACZEK, S., SCHOMBURG, I., CHANG, A., & SCHOMBURG, D. (2018): BRENDA in 2019: A European ELIXIR core data resource. *Nucleic acids research* (2018), vol. *47*([D1]): D542–D549. https://doi.org/10.1093/nar/gky1048 (cit. on pp. 13, 74, 95)

JONES, E., OLIPHANT, T., PETERSON, P., Et al. (2001): SciPy: Open source scientific tools for Python. http://www.scipy.org/. (Cit. on p. 183)

KANEHISA, M., FURUMICHI, M., TANABE, M., SATO, Y., & MORISHIMA, K. (2016): KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research* (2016), vol. *45*([D1]): D353–D361. https://doi.org/10.1093/nar/gkw1092 (cit. on pp. 60, 62)

KANEHISA, M., & GOTO, S. (2000): KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* (2000), vol. *28*([1]). https://doi.org/10.1093/nar/28.1.27 (cit. on pp. 39, 60)

KANEHISA, M., GOTO, S., SATO, Y., FURUMICHI, M., & TANABE, M. (2012): KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research* (2012), vol. *40*([D1]). https://doi.org/10.1093/nar/gkr988 (cit. on pp. 15, 16)

KANEHISA, M., SATO, Y., FURUMICHI, M., MORISHIMA, K., & TANABE, M. (2018): New approach for understanding genome variations in KEGG. *Nucleic acids research* (2018), vol. *47*([D1]): D590–D595. https://doi.org/10.1093/nar/gky962 (cit. on pp. 15, 16, 60)

KANEHISA, M., SATO, Y., KAWASHIMA, M., FURUMICHI, M., & TANABE, M. (2016): KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research* (2016), vol. *44*([D1]): D457–D462 (cit. on p. 39).

KARIM, A. S., & JEWETT, M. C. (2016): A cell-free framework for rapid biosynthetic pathway prototyping and enzyme discovery. *Metabolic engineering* (2016), vol. *36*: 116–126 (cit. on p. 37).

KARP, P. D., BILLINGTON, R., CASPI, R., FULCHER, C. A., LATENDRESSE, M., KOTHARI, A., KESELER, I. M., KRUMMENACKER, M., MIDFORD, P. E., ONG, Q., Et al. (2017): The BioCyc collection of microbial genomes and metabolic pathways. *Briefings in bioinformatics* (2017), vol. (cit. on pp. 15, 16).

KARP, P. D., LATENDRESSE, M., PALEY, S. M., KRUMMENACKER, M., ONG, Q. D., BILLINGTON, R., KOTHARI, A., WEAVER, D., LEE, T., SUBHRAVETI, P., Et al. (2015): Pathway Tools version 19.0 update: Software for pathway/genome informatics and systems biology. *Briefings in bioinformatics* (2015), vol. *17*([5]): 877–890 (cit. on pp. 18, 19).

KESELER, I. M., MACKIE, A., SANTOS-ZAVALETA, A., BILLINGTON, R., BONAVIDES-MARTÍNEZ, C., CASPI, R., FULCHER, C., GAMA-CASTRO, S., KOTHARI, A., KRUMMENACKER, M., LATENDRESSE, M., MUÑIZ-RASCADO, L., ONG, Q., PALEY, S., PERALTA-GIL, M., SUBHRAVETI, P., VELÁZQUEZ-RAMÍREZ, D. A., WEAVER, D., COLLADO-VIDES, J., ... KARP, P. D. (2017): The EcoCyc database: Reflecting new knowledge about *Escherichia coli* k-12. *Nucleic acids research* (D1 Jan. 2017), vol. *45*: D543–D550. https://doi.org/10.1093/nar/gkw1003 (cit. on pp. 14, 17, 19)

KHODAYARI, A., & MARANAS, C. D. (2016): A genome-scale *Escherichia coli* kinetic metabolic model k-ecoli457 satisfying flux data for multiple mutant strains. *Nature Communications* (2016), vol. *7* (cit. on p. 54).

KIM, S., CHEN, J., CHENG, T., GINDULYTE, A., HE, J., HE, S., LI, Q., SHOEMAKER, B. A., THIESSEN, P. A., YU, B., ZASLAVSKY, L., ZHANG, J., & BOLTON, E. E. (2019): Pubchem 2019 update: Improved access to chemical data. *Nucleic acids research* (D1 Jan. 2019), vol. *47*: D1102–D1109. https://doi.org/10.1093/nar/gky1033 (cit. on p. 11)

KIM, T. Y., SOHN, S. B., BIN KIM, Y., KIM, W. J., & LEE, S. Y. (2012): Recent advances in reconstruction and applications of genome-scale metabolic models. *Current Opinion in Biotechnology* (2012), vol. *23*([4]): 617–623. https://doi.org/10.1016/j.copbio.2011.10.007 (cit. on p. 29)

KING, Z. A., DRÄGER, A., EBRAHIM, A., SONNENSCHEIN, N., LEWIS, N. E., & PALSSON, B. O. (2015): Escher: A web application for building, sharing, and embedding data-rich visualizations of biological pathways. *PLoS computational biology* (2015), vol. *11*([8]): e1004321 (cit. on p. 15).

KING, Z. A., LU, J., DRÄGER, A., MILLER, P., FEDEROWICZ, S., LERMAN, J. A., EBRAHIM, A., PALSSON, B. O., & LEWIS, N. E. (2015): BiGG models: A platform for integrating,

standardizing and sharing genome-scale models. *Nucleic acids research* (2015), vol. *44*([D1]): D515–D522 (cit. on pp. 14, 15).

KLAMT, S., SAEZ-RODRIGUEZ, J., & GILLES, E. D. (2007): Structural and functional analysis of cellular networks with CellNetAnalyzer. *BMC Systems Biology* (2007), vol. *1*. https://doi.org/10.1186/1752-0509-1-2 (cit. on p. 30)

KOIZUMI, S. (2003): Large-scale production of oligosaccharides using bacterial functions. *Trends in Glycoscience and Glycotechnology* (2003), vol. *15*([82]): 65–74. https://doi.org/10.4052/tigg.15.65 (cit. on p. 37)

KOIZUMI, S., ENDO, T., TABATA, K., NAGANO, H., OHNISHI, J., & OZAKI, A. (2000): Large-scale production of GDP-fucose and Lewis X by bacterial coupling. *Journal of Industrial Microbiology & Biotechnology* (2000), vol. *25*([4]): 213–217. https://doi.org/10.1038/sj.jim.7000055 (cit. on p. 37)

KOIZUMI, S., ENDO, T., TABATA, K., & OZAKI, A. (1998): Large-scale production of UDP-galactose and globotriose by coupling metabolically engineered bacteria. *Nat Biotech* (1998), vol. *16*([9]): 847–850. http://dx.doi.org/10.1038/nbt0998-847 (cit. on p. 37)

KOTERA, M., HATTORI, M., OH, M.-A., YAMAMOTO, R., KOMENO, T., YABUZAKI, J., TONOMURA, K., GOTO, S., & KANEHISA, M. (2004): RPAIR: A reactant-pair database representing chemical changes in enzymatic reactions. *Genome Informatics* (2004), vol. *15*: P062 (cit. on pp. 39, 64).

KOTERA, M., OKUNO, Y., HATTORI, M., GOTO, S., & KANEHISA, M. (2004): Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. *Journal of the American Chemical Society* (2004), vol. *126*([50]): 16487–16498. https://doi.org/10.1021/ja0466457 (cit. on p. 64)

KRAUSER, S., HOFFMANN, T., & HEINZLE, E. (2015): Directed multistep biocatalysis for the synthesis of the polyketide oxytetracycline in permeabilized cells of *Escherichia coli*. *ACS Catalysis* (2015), vol. *5*([3]): 1407–1413. https://doi.org/10.1021/cs501825u (cit. on p. 37)

KRAUSER, S., KIEFER, P., & HEINZLE, E. (2012): Multienzyme whole-cell *in situ* biocatalysis for the production of flaviolin in permeabilized cells of *Escherichia coli*. *Chemcatchem* (2012), vol. *4*([6]). https://doi.org/10.1002/cctc.201100351 (cit. on p. 6)

KRIVORUCHKO, A., & NIELSEN, J. (2015): Production of natural products through metabolic engineering of *Saccharomyces cerevisiae*. *Current opinion in biotechnology* (2015), vol. *35*: 7–15. https://doi.org/10.1016/j.copbio.2014.12.004 (cit. on p. 70)

KROEMER, J. O., WITTMANN, C., SCHROEDER, H., & HEINZLE, E. (2006): Metabolic pathway analysis for rational design of l-methionine production by *Escherichia coli* and

Corynebacterium glutamicum. *Metabolic Engineering* (2006), vol. *8*([4]): 353–369. https://doi.org/10.1016/j.ymben.2006.02.001 (cit. on p. 10)

KUHN, D., KHOLIQ, M. A., HEINZLE, E., BUEHLER, B., & SCHMID, A. (2010): Intensification and economic and ecological assessment of a biocatalytic oxyfunctionalization process. *Green Chemistry* (2010), vol. *12*([5]): 815–827. https://doi.org/10.1039/b921896c (cit. on p. 4)

KUMAR, A., WANG, L., NG, C. Y., & MARANAS, C. D. (2018): Pathway design using *de novo* steps through uncharted biochemical spaces. *Nature communications* (1 Jan. 2018), vol. *9*: 184. https://doi.org/10.1038/s41467-017-02362-x (cit. on p. 24)

LALONDE, M.-E., & DUROCHER, Y. (2017): Therapeutic glycoprotein production in mammalian cells. *Journal of biotechnology* (2017), vol. *251*: 128–140. https://doi.org/10.1016/j.jbiotec.2017.04.028 (cit. on p. 69)

LE NOVERE, N., HUCKA, M., MI, H. Y., MOODIE, S., SCHREIBER, F., SOROKIN, A., DEMIR, E., WEGNER, K., ALADJEM, M. I., WIMALARATNE, S. M., BERGMAN, F. T., GAUGES, R., GHAZAL, P., KAWAJI, H., LI, L., MATSUOKA, Y., VILLEGER, A., BOYD, S. E., CALZONE, L., ... KITANO, H. (2009): The Systems Biology Graphical Notation. *Nature Biotechnology* (2009), vol. *27*([9]): 864–864. https://doi.org/10.1038/nbt0909-864d (cit. on p. 27)

LEE, W.-H., KIM, M.-D., JIN, Y.-S., & SEO, J.-H. (2013): Engineering of NADPH regenerators in *Escherichia coli* for enhanced biotransformation. *Applied Microbiology and Biotechnology* (2013), vol. *97*([7]): 2761–72. https://doi.org/10.1007/s00253-013-4750-z (cit. on p. 28)

LIAO, Y.-C., TSAI, M.-H., CHEN, F.-C., & HSIUNG, C. A. (2012): GEMSiRV: a software platform for GEnome-scale metabolic model simulation, reconstruction and visualization. *Bioinformatics* (2012), vol. *28*([13]): 1752–1758. https://doi.org/10.1093/bioinformatics/bts267 (cit. on p. 18)

LIESE, A., SEELBACH, K., & WANDREY, C. (2000): *Industrial biotransformations*. Wiley-VCH. (Cit. on p. 4).

LIN, G.-M., WARDEN-ROTHMAN, R., & VOIGT, C. A. (2019): Retrosynthetic design of metabolic pathways to chemicals not found in nature. *Current Opinion in Systems Biology* (2019), vol. https://doi.org/https://doi.org/10.1016/j.coisb.2019.04.004 (cit. on p. 80)

LOESCHCKE, A., & THIES, S. (2015): *Pseudomonas putida* — a versatile host for the production of natural products. *Applied microbiology and biotechnology* (2015), vol. *99*([15]): 6197–6214. https://doi.org/10.1007/s00253-015-6745-4 (cit. on p. 70)

LOPEZ-GALLEGO, F., & SCHMIDT-DANNERT, C. (2010): Multi-enzymatic synthesis. *Curr Opin Chem Biol* (2010), vol. *14*([2]): 174–83 (cit. on p. 3).

MA, H. W., & ZENG, A. P. (2003): The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics* (2003), vol. *19*([11]): 1423–30. http://www.ncbi.nlm.nih.gov/pubmed/12874056 (cit. on p. 6)

MAGLOTT, D., OSTELL, J., PRUITT, K. D., & TATUSOVA, T. (2011): Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research* (2011), vol. *39*: D52–D57. https://doi.org/10.1093/nar/gkq1237 (cit. on p. 12)

MAGNÚSDÓTTIR, S., & THIELE, I. (2018): Modeling metabolism of the human gut microbiome. *Current opinion in biotechnology* (June 2018), vol. *51*: 90–96. https://doi.org/10.1016/j.copbio.2017.12.005 (cit. on p. 79)

MCCLYMONT, K., & SOYER, O. S. (2013): Metabolic tinker: An online tool for guiding the design of synthetic metabolic pathways. *Nucleic acids research* (2013), vol. *41*([11]): e113–e113 (cit. on p. 37).

MI, H., SCHREIBER, F., MOODIE, S., CZAUDERNA, T., DEMIR, E., HAW, R., LUNA, A., LE NOVÈRE, N., SOROKIN, A., & VILLÉGER, A. (2015): Systems Biology Graphical Notation: Activity Flow language Level 1 Version 1.2. *Journal of integrative bioinformatics* (2 Sept. 2015), vol. *12*: 265. https://doi.org/10.2390/biecoll-jib-2015-265 (cit. on p. 28)

MICHAL, G., & SCHOMBURG, D. (2012): *Biochemical pathways. an atlas of biochemistry and molecular biology*. Wiley. (Cit. on pp. 16, 49).

MINTON, A. P. (2006): How can biochemical reactions within cells differ from those in test tubes? *Journal of Cell Science* (2006), vol. *119*([14]): 2863–2869. https://doi.org/Doi10.1242/Jcs.03063 (cit. on p. 5)

MONTI, D., FERRANDI, E. E., ZANELLATO, I., HUA, L., POLENTINI, F., CARREA, G., & RIVA, S. (2009): One-pot multienzymatic synthesis of 12-ketoursodeoxycholic acid: Subtle cofactor specificities rule the reaction equilibria of five biocatalysts working in a row. *Advanced Synthesis & Catalysis* (2009), vol. *351*([9]): 1303–1311. https://doi.org/10.1002/adsc.200800727 (cit. on p. 5)

MORETTI, S., MARTIN, O., VAN DU TRAN, T., BRIDGE, A., MORGAT, A., & PAGNI, M. (2016): MetaNetX/MNXref–reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. *Nucleic acids research* (D1 Jan. 2016), vol. *44*: D523–D526. https://doi.org/10.1093/nar/gkv1117 (cit. on p. 16)

MORGAT, A., LOMBARDOT, T., AXELSEN, K. B., AIMO, L., NIKNEJAD, A., HYKA-NOUSPIKEL, N., COUDERT, E., POZZATO, M., PAGNI, M., MORETTI, S., Et al. (2016): Updates in rhea - an expert curated resource of biochemical reactions. *Nucleic acids research* (2016), vol.: gkw990 (cit. on pp. 10, 11).

MOSES, T., POLLIER, J., THEVELEIN, J. M., & GOOSSENS, A. (2013): Bioengineering of plant (tri)terpenoids: From metabolic engineering of plants to synthetic biology *in vivo* and *in vitro*. *New Phytol* (2013), vol. *200*([1]): 27–43. https://doi.org/10.1111/nph.12325 (cit. on p. 5)

MUTO, A., KOTERA, M., TOKIMATSU, T., NAKAGAWA, Z., GOTO, S., & KANEHISA, M. (2013): Modular architecture of metabolic pathways revealed by conserved sequences of reactions. *Journal of Chemical Information and Modeling* (2013), vol. *53*([3]): 613–622. https://doi.org/10.1021/ci3005379 (cit. on p. 64)

NEUNER, A., & HEINZLE, E. (2011): Mixed glucose and lactate uptake by *Corynebacterium glutamicum* through metabolic engineering. *Biotechnology Journal* (2011), vol. *6*([3]): 318–329. https://doi.org/10.1002/biot.201000307 (cit. on pp. 10, 30)

NICOLAE, A., WAHRHEIT, J., NONNENMACHER, Y., WEYLER, C., & HEINZLE, E. (2015): Identification of active elementary flux modes in mitochondria using selectively permeabilized CHO cells. *Metab Eng* (2015), vol. *32*: 95–105. https://doi.org/10.1016/j.ymben.2015.09.014 (cit. on p. 65)

NIELSEN, J. (2019): Yeast systems biology: Model organism and cell factory. *Biotechnology Journal* (2019), vol. *0*([ja]): 1800421. https://doi.org/10.1002/biot.201800421 (cit. on p. 70)

NIKEL, P. I., CHAVARRIA, M., DANCHIN, A., & de LORENZO, V. (2016): From dirt to industrial applications: *Pseudomonas putida* as a synthetic biology chassis for hosting harsh biochemical reactions. *Current Opinion in Chemical Biology* (2016), vol. *34*: 20–29. https://doi.org/10.1016/j.cbpa.2016.05.011 (cit. on p. 70)

NOOR, E., BAR-EVEN, A., FLAMHOLZ, A., REZNIK, E., LIEBERMEISTER, W., & MILO, R. (2014): Pathway thermodynamics highlights kinetic obstacles in central metabolism. *PLoS computational biology* (2014), vol. *10*([2]): e1003483 (cit. on p. 23).

NOOR, E., HARALDSDÓTTIR, H. S., MILO, R., & FLEMING, R. M. (2013): Consistent estimation of gibbs energy using component contributions. *PLoS Comput Biol* (2013), vol. *9*([7]): e1003098. https://doi.org/10.1371/journal.pcbi.1003098 (cit. on pp. 23, 39, 40, 64)

OBERHARDT, M. A., PALSSON, B. O., & PAPIN, J. A. (2009): Applications of genome-scale metabolic reconstructions. *Molecular Systems Biology* (2009), vol. *5*. https://doi.org/10.1038/msb.2009.77 (cit. on p. 24)

ORCHARD, S., AMMARI, M., ARANDA, B., BREUZA, L., BRIGANTI, L., BROACKES-CARTER, F., CAMPBELL, N. H., CHAVALI, G., CHEN, C., DEL-TORO, N., DUESBURY, M., DUMOUSSEAU, M., GALEOTA, E., HINZ, U., IANNUCCELLI, M., JAGANNATHAN, S., JIMENEZ, R., KHADAKE, J., LAGREID, A., ... HERMJAKOB, H. (2014): The MIntAct project–IntAct as a common

curation platform for 11 molecular interaction databases. *Nucleic acids research* (Database issue Jan. 2014), vol. *42*: D358–D363. https://doi.org/10.1093/nar/gkt1115 (cit. on p. 13)

ORTH, J. D., CONRAD, T. M., NA, J., LERMAN, J. A., NAM, H., FEIST, A. M., & PALSSON, B. Ø. (2011): A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism - 2011. *Molecular systems biology* (2011), vol. *7*([1]). https://doi.org/10.1038/msb.2011.65 (cit. on p. 79)

OVERBEEK, R., BEGLEY, T., BUTLER, R. M., CHOUDHURI, J. V., CHUANG, H. Y., COHOON, M., de CRECY-LAGARD, V., DIAZ, N., DISZ, T., EDWARDS, R., FONSTEIN, M., FRANK, E. D., GERDES, S., GLASS, E. M., GOESMANN, A., HANSON, A., IWATA-REUYL, D., JENSEN, R., JAMSHIDI, N., ... VONSTEIN, V. (2005): The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research* (2005), vol. *33*([17]): 5691–5702. https://doi.org/10.1093/nar/gki866 (cit. on pp. 18, 19)

PEABODY, M. A., LAIRD, M. R., VLASSCHAERT, C., LO, R., & BRINKMAN, F. S. L. (2016): PSORTdb: Expanding the bacteria and archaea protein subcellular localization database to better reflect diversity in cell envelope structures. *Nucleic acids research* (D1 Jan. 2016), vol. *44*: D663–D668. https://doi.org/10.1093/nar/gkv1271 (cit. on p. 13)

PEIXOTO, T. P. (2014): The graph-tool python library. *figshare* (2014), vol. Retrieved Sept. 10, 2014, from http://figshare.com/articles/graph_tool/1164194 (cit. on pp. 67, 183)

PEY, J., PLANES, F. J., & BEASLEY, J. E. (2014): Refining carbon flux paths using atomic trace data. *Bioinformatics* (2014), vol. *30*([7]): 975 (cit. on p. 37).

PEY, J., PRADA, J., BEASLEY, J., & PLANES, F. (2011): Path finding methods accounting for stoichiometry in metabolic networks. *Genome biology* (May 2011), vol. *12*([5]): R49. http://www.ncbi.nlm.nih.gov/pubmed/21619601 (cit. on pp. 37, 39, 42–45, 61)

PHARKYA, P., BURGARD, A. P., & MARANAS, C. D. (2004): Optstrain: A computational framework for redesign of microbial production systems. *Genome research* (2004), vol. *14*([11]): 2367–2376 (cit. on p. 37).

POBLETE-CASTRO, I., BECKER, J., DOHNT, K., DOS SANTOS, V. M., & WITTMANN, C. (2012): Industrial biotechnology of *Pseudomonas putida* and related species. *Applied microbiology and biotechnology* (2012), vol. *93*([6]): 2279–2290. https://doi.org/10.1007/s00253-012-3928-0 (cit. on p. 70)

PONTRELLI, S., CHIU, T.-Y., LAN, E. I., CHEN, F. Y.-H., CHANG, P., & LIAO, J. C. (2018): *Escherichia coli* as a host for metabolic engineering. *Metabolic engineering* (2018), vol. *50*: 16–46. https://doi.org/10.1016/j.ymben.2018.04.008 (cit. on p. 70)

RALEVIC, V. (2015): Udp-glucose. *Reference module in biomedical sciences*. Elsevier. https://doi.org/10.1016/B978-0-12-801238-3.09699-9. (Cit. on p. 81)

REED, J. L., FAMILI, I., THIELE, I., & PALSSON, B. O. (2006): Towards multidimensional genome annotation. *Nature Reviews Genetics* (2006), vol. *7*([2]). https://doi.org/10.1038/nrg1769 (cit. on pp. 20, 30)

REITZ, M., SACHER, O., TARKHOV, A., TRÜMBACH, D., & GASTEIGER, J. (2004): Enabling the exploration of biochemical pathways. *Organic & biomolecular chemistry* (2004), vol. *2*([22]): 3226–3237 (cit. on p. 16).

ROUGNY, A., TOURÉ, V., MOODIE, S., BALAUR, I., CZAUDERNA, T., BORLINGHAUS, H., DOGRUSOZ, U., MAZEIN, A., DRÄGER, A., BLINOV, M. L., VILLÉGER, A., HAW, R., DEMIR, E., MI, H., SOROKIN, A., SCHREIBER, F., & LUNA, A. (2019): Systems Biology Graphical Notation: Process Description language Level 1 Version 2.0. *Journal of integrative bioinformatics* (2 June 2019), vol. *16*. https://doi.org/10.1515/jib-2019-0022 (cit. on p. 27)

RYAN, W., & PARULEKAR, S. J. (1991): Immobilization of *Escherichia coli* jm103 puc8 in kappa-carrageenan coupled with recombinant protein release by *in situ* cell-membrane permeabilization. *Biotechnology Progress* (1991), vol. *7*([2]): 99–110. https://doi.org/10.1021/bp00008a004 (cit. on p. 30)

SAIER JR, M. H., REDDY, V. S., TSU, B. V., AHMED, M. S., LI, C., & MORENO-HAGELSIEB, G. (2015): The transporter classification database (TCDB): recent advances. *Nucleic acids research* (2015), vol. *44*([D1]): D372–D379 (cit. on p. 11).

SASAKI, Y., ISHIKAWA, J., YAMASHITA, A., OSHIMA, K., KENRI, T., FURUYA, K., YOSHINO, C., HORINO, A., SHIBA, T., SASAKI, T., Et al. (2002): The complete genomic sequence of *Mycoplasma penetrans*, an intracellular bacterial pathogen in humans. *Nucleic acids research* (2002), vol. *30*([23]): 5293–5300. https://doi.org/10.1093/nar/gkf667 (cit. on p. 70)

SAYERS, E. W., BARRETT, T., BENSON, D. A., BOLTON, E., BRYANT, S. H., CANESE, K., CHETVERNIN, V., CHURCH, D. M., DICUCCIO, M., FEDERHEN, S., FEOLO, M., FINGERMAN, I. M., GEER, L. Y., HELMBERG, W., KAPUSTIN, Y., KRASNOV, S., LANDSMAN, D., LIPMAN, D. J., LU, Z. Y., ... YE, J. (2012): Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* (2012), vol. *40*([D1]): D13–D25. https://doi.org/10.1093/nar/gkr1184 (cit. on p. 12)

SCHELLENBERGER, J., QUE, R., FLEMING, R. M. T., THIELE, I., ORTH, J. D., FEIST, A. M., ZIELINSKI, D. C., BORDBAR, A., LEWIS, N. E., RAHMANIAN, S., KANG, J., HYDUKE, D. R., & PALSSON, B. O. (2011): Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox v2.0. *Nature Protocols* (2011), vol. *6*([9]): 1290–1307. https://doi.org/10.1038/nprot.2011.308 (cit. on p. 39)

SCHNEIDER, K., DORSCHEID, S., WITTE, K., GIFFHORN, F., & HEINZLE, E. (2012): Controlled feeding of hydrogen peroxide as oxygen source improves production of 5-ketofructose from L-sorbose using engineered pyranose 2-oxidase from *Peniophora gigantea*. *Biotechnology and Bioengineering* (2012), vol. *109*([11]): 2941–2945. https://doi.org/10.1002/bit.24572 (cit. on p. 4)

SCHUH, L. K., WEYLER, C., & HEINZLE, E. (2019): In-depth characterization of genome-scale network reconstructions for the in vitro synthesis in cell-free systems. *Biotechnology and Bioengineering* (2019), vol. *117*([4]): 1137–1147. https://doi.org/10.1002/bit.27249 (cit. on pp. 74, 76, 157, 172, 188)

SCHUSTER, S., DANDEKAR, T., & FELL, D. A. (1999): Detection of elementary flux modes in biochemical networks: A promising tool for pathway analysis and metabolic engineering. *Trends in biotechnology* (1999), vol. *17*([2]): 53–60 (cit. on p. 37).

SHANNON, P., MARKIEL, A., OZIER, O., BALIGA, N. S., WANG, J. T., RAMAGE, D., AMIN, N., SCHWIKOWSKI, B., & IDEKER, T. (2003): Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* (2003), vol. *13*([11]): 2498–2504 (cit. on p. 48).

SIEGEL, J. B., ZANGHELLINI, A., LOVICK, H. M., KISS, G., LAMBERT, A. R., CLAIR, J. L. S., GALLAHER, J. L., HILVERT, D., GELB, M. H., STODDARD, B. L., HOUK, K. N., MICHAEL, F. E., & BAKER, D. (2010): Computational design of an enzyme catalyst for a stereoselective bimolecular diels-alder reaction. *Science* (2010), vol. *329*([5989]): 309–313. https://doi.org/10.1126/science.1190239 (cit. on p. 3)

SMOOT, M. E., ONO, K., RUSCHEINSKI, J., WANG, P.-L., & IDEKER, T. (2011): Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* (2011), vol. *27*([3]): 431–432 (cit. on p. 48).

SOROKIN, A., LE NOVÈRE, N., LUNA, A., CZAUDERNA, T., DEMIR, E., HAW, R., MI, H., MOODIE, S., SCHREIBER, F., & VILLÉGER, A. (2015): Systems Biology Graphical Notation: Entity Relationship language Level 1 Version 2. *Journal of integrative bioinformatics* (2 Sept. 2015), vol. *12*: 264. https://doi.org/10.2390/biecoll-jib-2015-264 (cit. on p. 28)

SOROKINA, M., STAM, M., MÉDIGUE, C., LESPINET, O., & VALLENET, D. (2014): Profiling the orphan enzymes. *Biology direct* (2014), vol. *9*([1]): 10 (cit. on p. 93).

SRINIVASAN, S., CLUETT, W. R., & MAHADEVAN, R. (2015): Constructing kinetic models of metabolism at genome-scales: A review. *Biotechnology journal* (2015), vol. *10*([9]): 1345–1359 (cit. on p. 94).

STEENSELS, J., SNOEK, T., MEERSMAN, E., NICOLINO, M. P., VOORDECKERS, K., & VERSTREPEN, K. J. (2014): Improving industrial yeast strains: Exploiting natural and artificial

diversity. *FEMS microbiology reviews* (2014), vol. *38*([5]): 947–995. https://doi.org/10.1111/1574-6976.12073 (cit. on p. 70)

SZKLARCZYK, D., GABLE, A. L., LYON, D., JUNGE, A., WYDER, S., HUERTA-CEPAS, J., SIMONOVIC, M., DONCHEVA, N. T., MORRIS, J. H., BORK, P., JENSEN, L. J., & MERING, C. V. (2019): String v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research* (D1 Jan. 2019), vol. *47*: D607–D613. https://doi.org/10.1093/nar/gky1131 (cit. on p. 13)

TAKEGAWA, K., TOHDA, H., SASAKI, M., IDIRIS, A., OHASHI, T., MUKAIYAMA, H., GIGA-HAMA, Y., & KUMAGAI, H. (2009): Production of heterologous proteins using the fission-yeast (*Schizosaccharomyces pombe*) expression system. *Biotechnology and applied biochemistry* (2009), vol. *53*([4]): 227–235. https://doi.org/10.1042/BA20090048 (cit. on p. 70)

TERVO, C. J., & REED, J. L. (2016): MapMaker and PathTracer for tracking carbon in genome-scale metabolic models. *Biotechnology journal* (2016), vol. *11*([5]): 648–661 (cit. on p. 37).

TERZER, M., MAYNARD, N. D., COVERT, M. W., & STELLING, J. (2009): Genome-scale metabolic networks. *Wiley Interdisciplinary Reviews-Systems Biology and Medicine* (2009), vol. *1*([3]). https://doi.org/10.1002/wsbm.37 (cit. on p. 29)

TERZER, M., & STELLING, J. (2008): Large-scale computation of elementary flux modes with bit pattern trees. *Bioinformatics* (2008), vol. *24*([19]): 2229–2235. https://doi.org/10.1093/bioinformatics/btn401 (cit. on p. 18)

THIELE, I., & PALSSON, B. O. (2010): A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols* (2010), vol. *5*([1]). https://doi.org/10.1038/nprot.2009.203 (cit. on pp. 20, 22–24)

von KAMP, A., & SCHUSTER, S. (2006): Metatool 5.0: Fast and flexible elementary modes analysis. *Bioinformatics* (2006), vol. *22*([15]): 1930–1931. https://doi.org/10.1093/bioinformatics/btl267 (cit. on pp. 18, 19, 30)

von KAMP, A., THIELE, S., HÄDICKE, O., & KLAMT, S. (2017): Use of cellnetanalyzer in biotechnology and metabolic engineering. *Journal of biotechnology* (2017), vol. *261*: 221–228. https://doi.org/10.1016/j.jbiotec.2017.05.001 (cit. on pp. 18, 19, 79)

WANG, L., DASH, S., NG, C. Y., & MARANAS, C. D. (2017): A review of computational tools for design and reconstruction of metabolic pathways. *Synthetic and systems biotechnology* (2017), vol. *2*([4]): 243–252. https://doi.org/10.1016/j.synbio.2017.11.002 (cit. on p. 60)

WANG, S.-Z., ZHANG, Y.-H., REN, H., WANG, Y.-L., JIANG, W., & FANG, B.-S. (2017): Strategies and perspectives of assembling multi-enzyme systems. *Critical reviews in biotechnology* (2017), vol. *37*([8]): 1024–1037. https://doi.org/10.1080/07388551.2017.1303803 (cit. on pp. 60, 79)

WEYLER, C., & HEINZLE, E. (2015): Multistep synthesis of UDP-glucose using tailored, permeabilized cells of *E. coli*. *Applied Biochemistry and Biotechnology* (2015), vol. *175*([8]): 3729–3736. https://doi.org/10.1007/s12010-015-1540-3 (cit. on pp. 37, 81, 90)

WEYLER, C., & HEINZLE, E. (2017): Synthesis of natural variants and synthetic derivatives of the cyclic nonribosomal peptide luminmide in permeabilized *E. coli Nissle* and product formation kinetics. *Applied Microbiology and Biotechnology* (2017), vol. *101*([1]): 131–138 (cit. on p. 65).

WITTIG, U., KANIA, R., GOLEBIEWSKI, M., REY, M., SHI, L., JONG, L., ALGAA, E., WEIDEMANN, A., SAUER-DANZWITH, H., MIR, S., KREBS, O., BITTKOWSKI, M., WETSCH, E., ROJAS, I., & MUELLER, W. (2012): SABIO-RK-database for biochemical reaction kinetics. *Nucleic Acids Research* (2012), vol. *40*([D1]): D790–D796. https://doi.org/10.1093/nar/gkr1046 (cit. on p. 11)

WOHLGEMUTH, R. (2010): Biocatalysis - key to sustainable industrial chemistry. *Current Opinion in Biotechnology* (2010), vol. *21*([6]): 713–724. https://doi.org/10.1016/j.copbio.2010.09.016 (cit. on p. 4)

WOHLGEMUTH, R. (2011): Molecular and engineering perspectives of the biocatalysis interface to chemical synthesis. *Chemical and Biochemical Engineering Quarterly* (2011), vol. *25*([1]): 125–134 (cit. on p. 4).

WRÓTNIAK-DRZEWIECKA, W., BRZEZISKA, A. J., DAHM, H., INGLE, A. P., & RAI, M. (2016): Current trends in myxobacteria research. *Annals of microbiology* (2016), vol. *66*([1]): 17–33. https://doi.org/10.1007/s13213-015-1104-3 (cit. on p. 70)

WURM, F. M. (2004): Production of recombinant protein therapeutics in cultivated mammalian cells. *Nature biotechnology* (2004), vol. *22*([11]): 1393. https://doi.org/10.1038/nbt1026 (cit. on p. 69)

XIE, C., MAO, X., HUANG, J., DING, Y., WU, J., DONG, S., KONG, L., GAO, G., LI, C.-Y., & WEI, L. (2011): KOBAS 2.0: A web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Research* (2011), vol. *39*. https://doi.org/10.1093/nar/gkr483 (cit. on p. 18)

YE, X., HONDA, K., SAKAI, T., OKANO, K., OMASA, T., HIROTA, R., KURODA, A., & OHTAKE, H. (2012): Synthetic metabolic engineering-a novel, simple technology for designing a

chimeric metabolic pathway. *Microbial Cell Factories* (2012), vol. *11*([120]). https://doi.org/10.1186/1475-2859-11-120 (cit. on p. 6)

YOU, C., & ZHANG, Y. H. (2013): Cell-free biosystems for biomanufacturing. *Adv Biochem Eng Biotechnol* (2013), vol. *131*: 89–119. https://doi.org/10.1007/10_2012_159 (cit. on pp. 4, 5)

YU, N. Y., LAIRD, M. R., SPENCER, C., & BRINKMAN, F. S. L. (2011): PSORTdb - an expanded, auto-updated, user-friendly protein subcellular localization database for bacteria and archaea. *Nucleic Acids Research* (2011), vol. *39*. https://doi.org/10.1093/nar/gkq1093 (cit. on p. 13)

ZOMORRODI, A. R., & MARANAS, C. D. (2010): Improving the iMM904 *S. cerevisiae* metabolic model using essentiality and synthetic lethality data. *BMC Systems Biology* (Dec. 2010), vol. *4*: 178. https://doi.org/10.1186/1752-0509-4-178 (cit. on p. 79)

ZOMORRODI, A. R., SUTHERS, P. F., RANGANATHAN, S., & MARANAS, C. D. (2012): Mathematical optimization applications in metabolic networks. *Metabolic Engineering* (2012), vol. *14*([6]). https://doi.org/10.1016/j.ymben.2012.09.005 (cit. on pp. 20, 30)

List of Figures

List of Tables

Acronyms

$FADH_2$ reduced flavin adenine dinucleotide

NAD^+ nicotinamide adenine dinucleotide

ADP adenosine 5'-diphosphate

AF Activity Flow

AMP adenosine 5'-monophosphate

ATP adenosine 5'-triphosphate

BFS breadth-first search

BioPax Biological Pathway Exchange

BLAST Basic Local Alignment Search Tool

BRENDA Braunschweig Enzyme Database

ChEBI Chemical Entities of Biological Interest

CMP Cytidine-5'-monophosphate

COBRA COnstraint-Based Reconstruction and Analysis

CTP Cytidine 5'-triphosphate

DMAPP dimethylallyl pyrophosphate

DNA deoxyribonucleic acid

EC Enzyme Commission

ER Entity Relationship

GLAMM Genome-linked Application for Metabolic Maps

GO Gene Ontology

GPP geranyl pyrophosphate

IPP isopentenyl pyrophosphate

KEGG Kyoto Encyclopedia of Genes and Genomes

MILP mixed-integer linear program

MINE Metabolic *in silico* Network Expansions

NAD(P)H nicotinamide adenine dinucleotide (phosphate)

NADH reduced nicotinamide adenine dinucleotide

NCBI National Center for Biotechnology Information

PD Process Description

PGDB pathway/genome database

SBGN Systems Biology Graphical Notation

SBML Systems Biology Markup Language

TC transporter classification

TCDB Transporter Classification Database

UDP uridine 5'-diphosphate

UMP uridine 5'-monophosphate

UTP uridine 5'-triphosphate

Glossary

K_m Michaelis-Menten constant

pH potential of Hydrogen

°C degree Celcius

de novo from the beginning, anew

in silico refers to an experiment performed with a computer (simulation)

in situ refers to an experiment or a synthesis performed in place where it occurs

in vitro refers to a biological experiment performed outside the normal biological context

in vivo refers to an experiment or a synthesis performed in whole, living organisms or cells

MATLAB MATrix LABoratory; Programming language and numerical computation environment by MathWorks.

Python A high-level interpreted and object-oriented programming language

A Appendix Network Design and Analysis for Multi-Enzyme Biocatalysis

The following chapter is based on the the supplementary material of the research article (BLASS et al., 2017) (Chapter 4 Network Design and Analysis for Multi-Enzyme Biocatalysis).

A.1 Pathway Examples

Geranyl Pyrophosphate

Pathway Candidate: Mevalonate Pathway

Table A.1: Ranking of pathway candidate to geranyl pyrophosphate: mevalonate pathway

criterion	value
number of active reactions:	8
starts with basic:	False
reactions w/o dG:	0
sum (dG + |dG|):	0
dG:	-2.154517e+02
number of heterologous enzymes:	5
number of cofactors:	8
number of side reactions:	11

```
R08549 : 2-Oxoglutarate dehydrogenase complex -2.723933e+01

  substrates:
    1 C00026  2-Oxoglutarate
    1 C00010  CoA
    1 C00003  NAD+

  products:
```

```
    1 C00091  Succinyl-CoA
    1 C00011  CO2
    1 C00004  NADH
    1 C00080  H+

R02084 : succinyl-CoA:3-hydroxy-3-methylglutarate CoA-transferase -3.166626e+00

  substrates:
    1 C00091  Succinyl-CoA
    1 C03761  3-Hydroxy-3-methylglutarate

  products:
    1 C00042  Succinate
    1 C00356  (S)-3-Hydroxy-3-methylglutaryl-CoA

R02082 : (R)-Mevalonate:NADP+ oxidoreductase (CoA acylating)  -2.534910e+01

  substrates:
    1 C00356  (S)-3-Hydroxy-3-methylglutaryl-CoA
    2 C00005  NADPH
    2 C00080  H+

  products:
    1 C00418  (R)-Mevalonate
    1 C00010  CoA
    2 C00006  NADP+

R02245 : ATP:(R)-mevalonate 5-phosphotransferase  -1.348146e+01

  substrates:0
    1 C00002  ATP
    1 C00418  (R)-Mevalonate

  products:
    1 C00008  ADP
    1 C01107  (R)-5-Phosphomevalonate

R03245 : ATP:(R)-5-phosphomevalonate phosphotransferase -2.712405e+00

  substrates:
    1 C00002  ATP
    1 C01107  (R)-5-Phosphomevalonate

  products:
    1 C00008  ADP
    1 C01143  (R)-5-Diphosphomevalonate

R01121 : ATP:(R)-5-diphosphomevalonate carboxy-lyase (adding ATP  -6.741552e+01
```

```
  substrates:
    1 C00002  ATP
    1 C01143  (R)-5-Diphosphomevalonate

  products:
    1 C00008  ADP
    1 C00009  Orthophosphate
    1 C00129  Isopentenyl diphosphate
    1 C00011  CO2

R01123 : Isopentenyl-diphosphate delta3-delta2-isomerase  -4.902559e+00

  substrates:
    1 C00129  Isopentenyl diphosphate

  products:
    1 C00235  Dimethylallyl diphosphate

R01658 : Dimethylallyl-diphosphate:isopentenyl-diphosphate
dimethylallyltranstransferase  -7.118467e+01

  substrates:
    1 C00235  Dimethylallyl diphosphate
    1 C00129  Isopentenyl diphosphate

  products:
    1 C00013  Diphosphate
    1 C00341  Geranyl diphosphate
```

Overall Balance

Table A.2 shows the overall balance of the pathway candidate representing the mevalonate pathway.

Side Reactions

```
R00089 : ATP diphosphate-lyase (cyclizing; 3',5'-cyclic-AMP-forming)

  substrates:
    1 C00002  ATP

  products:
    1 C00575  3',5'-Cyclic AMP
    1 C00013  Diphosphate
```

Table A.2: Overall balance of the pathway candidate representing the mevalonate pathway

	R01121	R01123	R01658	R02082	R02084	R02245	R03245	R08549	overall
ATP	-1	0	0	0	0	-1	-1	0	-3
(R)-5-diphospho-mevalonate	-1	0	0	0	0	0	1	0	0
ADP	1	0	0	0	0	1	1	0	3
orthophosphate	1	0	0	0	0	0	0	0	1
isopentenyl diphosphate	1	-1	-1	0	0	0	0	0	-1
CO_2	1	0	0	0	0	0	0	1	2
dimethylallyl diphosphate	0	1	-1	0	0	0	0	0	0
diphosphate	0	0	1	0	0	0	0	0	1
geranyl diphosphate	0	0	1	0	0	0	0	0	1
(S)-3-hydroxy-3-methylglutaryl-CoA	0	0	0	-1	1	0	0	0	0
NADPH	0	0	0	-2	0	0	0	0	-2
H^+	0	0	0	-2	0	0	0	1	-1
(R)-mevalonate	0	0	0	1	0	-1	0	0	0
CoA	0	0	0	1	0	0	0	-1	0
$NADP^+$	0	0	0	2	0	0	0	0	2
succinyl-CoA	0	0	0	0	-1	0	0	1	0
3-hydroxy-3-methylglutarate	0	0	0	0	-1	0	0	0	-1
succinate	0	0	0	0	1	0	0	0	1
(R)-5-phosphomevalonate	0	0	0	0	0	1	-1	0	0
2-oxoglutarate	0	0	0	0	0	0	0	-1	-1
NAD^+	0	0	0	0	0	0	0	-1	-1
NADH	0	0	0	0	0	0	0	1	1

```
R00104 : ATP:NAD+ 2'-phosphotransferase

  substrates:
    1 C00002  ATP
    1 C00003  NAD+

  products:
    1 C00008  ADP
    1 C00006  NADP+

R00112 : NADPH:NAD+ oxidoreductase

  substrates:
    1 C00005  NADPH
    1 C00003  NAD+

  products:
    1 C00006  NADP+
    1 C00004  NADH

R00405 : Succinate:CoA ligase (ADP-forming)

  substrates:
    1 C00002  ATP
    1 C00042  Succinate
    1 C00010  CoA

  products:
    1 C00008  ADP
    1 C00009  Orthophosphate
    1 C00091  Succinyl-CoA

R02003 : Geranyl-diphosphate:isopentenyl-diphosphate geranyltrans-transferase

  substrates:
    1 C00341  Geranyl diphosphate
    1 C00129  Isopentenyl diphosphate

  products:
    1 C00013  Diphosphate
    1 C00448  trans,trans-Farnesyl diphosphate

-R00127 : -ATP:AMP phosphotransferase

  substrates:
    2 C00008  ADP
```

```
  products:
    1 C00002  ATP
    1 C00020  AMP

-R00130 : -ATP:dephospho-CoA 3'-phosphotransferase

  substrates:
    1 C00008  ADP
    1 C00010  CoA

  products:
    1 C00002  ATP
    1 C00882  Dephospho-CoA

-R00137 : -ATP:nicotinamide-nucleotide adenylyltransferase

  substrates:
    1 C00013  Diphosphate
    1 C00003  NAD+

  products:
    1 C00002  ATP
    1 C00455  Nicotinamide D-ribonucleotide

-R00267 : -Isocitrate:NADP+ oxidoreductase (decarboxylating)

  substrates:
    1 C00026  2-Oxoglutarate
    1 C00011  CO2
    1 C00005  NADPH
    1 C00080  H+

  products:
    1 C00311  Isocitrate
    1 C00006  NADP+

-R00519 : -formate:NAD+ oxidoreductase

  substrates:
    1 C00080  H+
    1 C00011  CO2
    1 C00004  NADH

  products:
    1 C00058  Formate
    1 C00003  NAD+
```

```
-R00833 : -(R)-Methylmalonyl-CoA CoA-carbonylmutase

  substrates:
    1 C00091  Succinyl-CoA

  products:
    1 C01213  (R)-Methylmalonyl-CoA
```

Pathway Candidate: Non-mevalonate Pathway

Table A.3: Ranking of pathway candidate to geranyl pyrophosphate: non-mevalonate pathway

criterion	value
number of active reactions:	9
starts with basic:	False
reactions w/o dG:	2
sum (dG + |dG|):	0
dG:	-1.998150e+02
number of heterologous enzymes:	0
number of cofactors:	3
number of side reactions:	24

```
R05636 : 1-Deoxy-D-xylulose-5-phosphate pyruvate-lyase (carboxylating)  -2.901798
e+01

  substrates:
    1 C00022  Pyruvate
    1 C00118  D-Glyceraldehyde 3-phosphate

  products:
    1 C11437  1-Deoxy-D-xylulose 5-phosphate
    1 C00011  CO2

R05688 : 1-Deoxy-D-xylulose-5-phosphate isomeroreductase  -2.321112e+01

  substrates:
    1 C11437  1-Deoxy-D-xylulose 5-phosphate
    1 C00005  NADPH
    1 C00080  H+

  products:
    1 C11434  2-C-Methyl-D-erythritol 4-phosphate
    1 C00006  NADP+
```

```
R05633 : CTP: 2-C-Methyl-D-erythritol 4-phosphate cytidylyltransferase  -4.729713
e+00

  substrates:
    1 C11434  2-C-Methyl-D-erythritol 4-phosphate
    1 C00063  CTP

  products:
    1 C11435  4-(Cytidine 5'-diphospho)-2-C-methyl-D-erythritol
    1 C00013  Diphosphate

R05634 : ATP:4-(Cytidine 5'-diphospho)-2-C-methyl-D-erythritol 2-
phosphotransferase -5.273891e+00

  substrates:
    1 C11435  4-(Cytidine 5'-diphospho)-2-C-methyl-D-erythritol
    1 C00002  ATP

  products:
    1 C11436  2-Phospho-4-(cytidine 5'-diphospho)-2-C-methyl-D-erythritol
    1 C00008  ADP

R05637 : 2-Phospho-4-(cytidine 5'-diphospho)-2-C-methyl-D-erythritol CMP-lyase (
cyclizing)  0

  substrates:
    1 C11436  2-Phospho-4-(cytidine 5'-diphospho)-2-C-methyl-D-erythritol

  products:
    1 C11453  2-C-Methyl-D-erythritol 2,4-cyclodiphosphate
    1 C00055  CMP

R08689 : (E)-4-hydroxy-3-methylbut-2-en-1-yl-diphosphate:oxidized ferredoxin
oxidoreductase (hydrating) 0

  substrates:
    1 C11453  2-C-Methyl-D-erythritol 2,4-cyclodiphosphate
    2 C00138  Reduced ferredoxin

  products:
    1 C11811  1-Hydroxy-2-methyl-2-butenyl 4-diphosphate
    1 C00001  H2O
    2 C00139  Oxidized ferredoxin

R08210 : dimethylallyl diphosphate:ferredoxin oxidoreductase  -6.639758e+01

  substrates:
    1 C11811  1-Hydroxy-2-methyl-2-butenyl 4-diphosphate
```

```
    2 C00138  Reduced ferredoxin
    2 C00080  H+

  products:
    1 C00235  Dimethylallyl diphosphate
    1 C00001  Oxidized ferredoxin
    2 C00139  H2O

R01658 : Dimethylallyl-diphosphate:isopentenyl-diphosphate
dimethylallyltranstransferase  -7.118467e+01

  substrates:
    1 C00235  Dimethylallyl diphosphate
    1 C00129  Isopentenyl diphosphate

  products:
    1 C00013  Diphosphate
    1 C00341  Geranyl diphosphate
----------
-R01123 : -Isopentenyl-diphosphate delta3-delta2-isomerase  0

  substrates:
    1 C00235  Dimethylallyl diphosphate

  products:
    1 C00129  Isopentenyl diphosphate
```

Overall Balance

Table A.4 shows the overall balance of the pathway candidate representing the non-mevalonate pathway.

Side Reactions

```
R00004 : diphosphate phosphohydrolase

  substrates:
    1 C00013  Diphosphate
    1 C00001  H2O

  products:
    2 C00009  Orthophosphate

R00086 : ATP phosphohydrolase
```

Table A.4: Overall balance of the pathway candidate representing the non-mevalonate pathway

	R01658	R05633	R05634	R05636	R05637	R05688	R08210	R08689	-R01123	overall
dimethylallyl diphosphate	-1	0	0	0	0	0	1	0	-1	-1
isopentenyl diphosphate	-1	0	0	0	0	0	0	0	1	0
diphosphate	1	1	0	0	0	0	0	0	0	2
geranyl diphosphate	1	0	0	0	0	0	0	0	0	1
2-C-methyl-D-erythritol 4-phosphate	0	-1	0	0	0	1	0	0	0	0
CTP	0	-1	0	0	0	0	0	0	0	-1
4-(cytidine 5-diphospho)-2-C-methyl-D-erythritol	0	1	-1	0	0	0	0	0	0	0
ATP	0	0	-1	0	0	0	0	0	0	-1
2-phospho-4-(cytidine 5-diphospho)-2-C-methyl-D-erythritol	0	0	1	0	-1	0	0	0	0	0
ADP	0	0	1	0	0	0	0	0	0	1
pyruvate	0	0	0	-1	0	0	0	0	0	-1
D-glyceraldehyde 3-phosphate	0	0	0	-1	0	0	0	0	0	-1

Table A.4: Overall balance of the pathway candidate representing the non-mevalonate pathway (continued)

	R01658	R05633	R05634	R05636	R05637	R05688	R08210	R08689	-R01123	overall
1-deoxy-D-xylulose 5-phosphate	0	0	0	1	0	-1	0	0	0	0
CO_2	0	0	0	1	0	0	0	0	0	1
2-C-methyl-D-erythritol 2,4-cyclodiphosphate	0	0	0	0	1	0	0	-1	0	0
CMP	0	0	0	0	1	0	0	0	0	1
NADPH	0	0	0	0	0	-1	0	0	0	-1
H^+	0	0	0	0	0	-1	-2	0	0	-3
NADP+	0	0	0	0	0	1	0	0	0	1
1-hydroxy-2-methyl-2-butenyl 4-diphosphate	0	0	0	0	0	0	-1	1	0	0
reduced ferredoxin	0	0	0	0	0	0	-2	-2	0	-4
oxidized ferredoxin	0	0	0	0	0	0	2	2	0	4
H_2O	0	0	0	0	0	0	1	1	0	2

```
  substrates:
    1 C00002  ATP
    1 C00001  H2O

  products:
    1 C00008  ADP
    1 C00009  Orthophosphate

R00087 : ATP diphosphohydrolase (diphosphate-forming)

  substrates:
    1 C00002  ATP
    1 C00001  H2O

  products:
    1 C00020  AMP
    1 C00013  Diphosphate

R00089 : ATP diphosphate-lyase (cyclizing

  substrates:
    1 C00002  ATP

  products:
    1 C00575  3,5-Cyclic AMP
    1 C00013  Diphosphate

R00199 : ATP:pyruvate,water phosphotransferase

  substrates:
    1 C00002  ATP
    1 C00022  Pyruvate
    1 C00001  H2O

  products:
    1 C00020  AMP
    1 C00074  Phosphoenolpyruvate
    1 C00009  Orthophosphate

R00200 : ATP:pyruvate 2-O-phosphotransferase

  substrates:
    1 C00002  ATP
    1 C00022  Pyruvate

  products:
    1 C00008  ADP
    1 C00074  Phosphoenolpyruvate
```

```
R00511 : cytidine-5-monophosphate phosphohydrolase

  substrates:
    1 C00055  CMP
    1 C00001  H2O

  products:
    1 C00475  Cytidine
    1 C00009  Orthophosphate

R00512 : ATP:CMP phosphotransferase

  substrates:
    1 C00002  ATP
    1 C00055  CMP

  products:
    1 C00008  ADP
    1 C00112  CDP

R00515 : CTP diphosphohydrolase (diphosphate-forming)

  substrates:
    1 C00063  CTP
    1 C00001  H2O

  products:
    1 C00055  CMP
    1 C00013  Diphosphate

R00568 : CTP aminohydrolase

  substrates:
    1 C00063  CTP
    1 C00001  H2O

  products:
    1 C00075  UTP
    1 C00014  Ammonia

R00572 : CTP:pyruvate 2-O-phosphotransferase

  substrates:
    1 C00063  CTP
    1 C00022  Pyruvate

  products:
```

```
    1 C00112  CDP
    1 C00074  Phosphoenolpyruvate

R01015 : D-glyceraldehyde-3-phosphate aldose-ketose-isomerase

  substrates:
    1 C00118  D-Glyceraldehyde 3-phosphate

  products:
    1 C00111  Glycerone phosphate

R01123 : Isopentenyl-diphosphate delta3-delta2-isomerase

  substrates:
    1 C00129  Isopentenyl diphosphate

  products:
    1 C00235  Dimethylallyl diphosphate

R01195 : Ferredoxin:NADP+ oxidoreductase

  substrates:
    1 C00006  Reduced ferredoxin
    1 C00080  NADP+
    2 C00138  H+

  products:
    1 C00005  Oxidized ferredoxin
    2 C00139  NADPH

R02003 : Geranyl-diphosphate:isopentenyl-diphosphate geranyltrans-transferase

  substrates:
    1 C00341  Geranyl diphosphate
    1 C00129  Isopentenyl diphosphate

  products:
    1 C00013  Diphosphate
    1 C00448  trans,trans-Farnesyl diphosphate

R05884 : isopentenyl-diphosphate:ferredoxin oxidoreductase

  substrates:
    1 C11811  1-Hydroxy-2-methyl-2-butenyl 4-diphosphate
    2 C00138  Reduced ferredoxin
    2 C00080  H+

  products:
```

```
    1 C00129  Isopentenyl diphosphate
    1 C00001  Oxidized ferredoxin
    2 C00139  H2O

-R00104 : -ATP:NAD+ 2-phosphotransferase

  substrates:
    1 C00008  ADP
    1 C00006  NADP+

  products:
    1 C00002  ATP
    1 C00003  NAD+

-R00127 : -ATP:AMP phosphotransferase

  substrates:
    2 C00008  ADP

  products:
    1 C00002  ATP
    1 C00020  AMP

-R00132 : -carbonate hydro-lyase (carbon-dioxide-forming)

  substrates:
    1 C00011  CO2
    1 C00001  H2O

  products:
    1 C01353  Carbonic acid

-R00216 : -(S)-Malate:NADP+ oxidoreductase(oxaloacetate-decarboxylating)

  substrates:
    1 C00022  Pyruvate
    1 C00011  CO2
    1 C00005  NADPH
    1 C00080  H+

  products:
    1 C00149  (S)-Malate
    1 C00006  NADP+

-R00513 : -ATP:cytidine 5-phosphotransferase

  substrates:
    1 C00008  ADP
```

```
    1 C00055  CMP

  products:
    1 C00002  ATP
    1 C00475  Cytidine

-R01064 : -2-dehydro-3-deoxy-6-phospho-D-galactonate D-glyceraldehyde-3-phospho-
lyase (pyruvate-forming)

  substrates:
    1 C00022  Pyruvate
    1 C00118  D-Glyceraldehyde 3-phosphate

  products:
    1 C01286  2-Dehydro-3-deoxy-6-phospho-D-galactonate

-R05605 : -2-dehydro-3-deoxy-6-phospho-D-gluconate D-glyceraldehyde-3-phosphate-
lyase (pyruvate-forming)

  substrates:
    1 C00118  D-Glyceraldehyde 3-phosphate
    1 C00022  Pyruvate

  products:
    1 C04442  2-Dehydro-3-deoxy-6-phospho-D-gluconate

-R10092 : -carbonate hydro-lyase (carbon-dioxide-forming)

  substrates:
    1 C00011  CO2
    1 C00001  H2O

  products:
    1 C00288  HCO3-
    1 C00080  H+
```

Amygdalin

Pathway Candidate: Starting From Sucrose

```
R00803 : sucrose:phosphate alpha-D-glucosyltransferase  4.881509e-02

  substrates:
    1 C00089  Sucrose
    1 C00009  Orthophosphate

  products:
```

Table A.5: Ranking of pathway candidate to amygdalin starting from sucrose

criterion	value
number of active reactions:	4
starts with basic:	True
reactions w/o dG:	0
sum (dG + |dG|):	9.763018e-02
dG:	-4.130443e+01
number of heterologous enzymes:	2
number of cofactors:	1
number of side reactions:	4

```
    1 C00095  D-Fructose
    1 C00103  D-Glucose 1-phosphate

R00289 : UTP:alpha-D-glucose-1-phosphate uridylyltransferase  -8.295754e+00

  substrates:
    1 C00075  UTP
    1 C00103  D-Glucose 1-phosphate

  products:
    1 C00013  Diphosphate
    1 C00029  UDP-glucose

R10638 : UDP-D-glucose:(R)-mandelonitrile beta-D-glucosyltransferase  -1.412458e
+01

  substrates:
    1 C00561  Mandelonitrile
    1 C00029  UDP-glucose

  products:
    1 C00844  Prunasin
    1 C00015  UDP

R10639 :  -1.893291e+01

  substrates:
    1 C00844  Prunasin
    1 C00029  UDP-glucose

  products:
    1 C08325  Amygdalin
    1 C00015  UDP
```

Thermodynamic Profile

Figure A.1 shows the thermodynamic profile for the pathway candidate for the synthesis of amygdalin from α-D-glucose 6-phosphate.

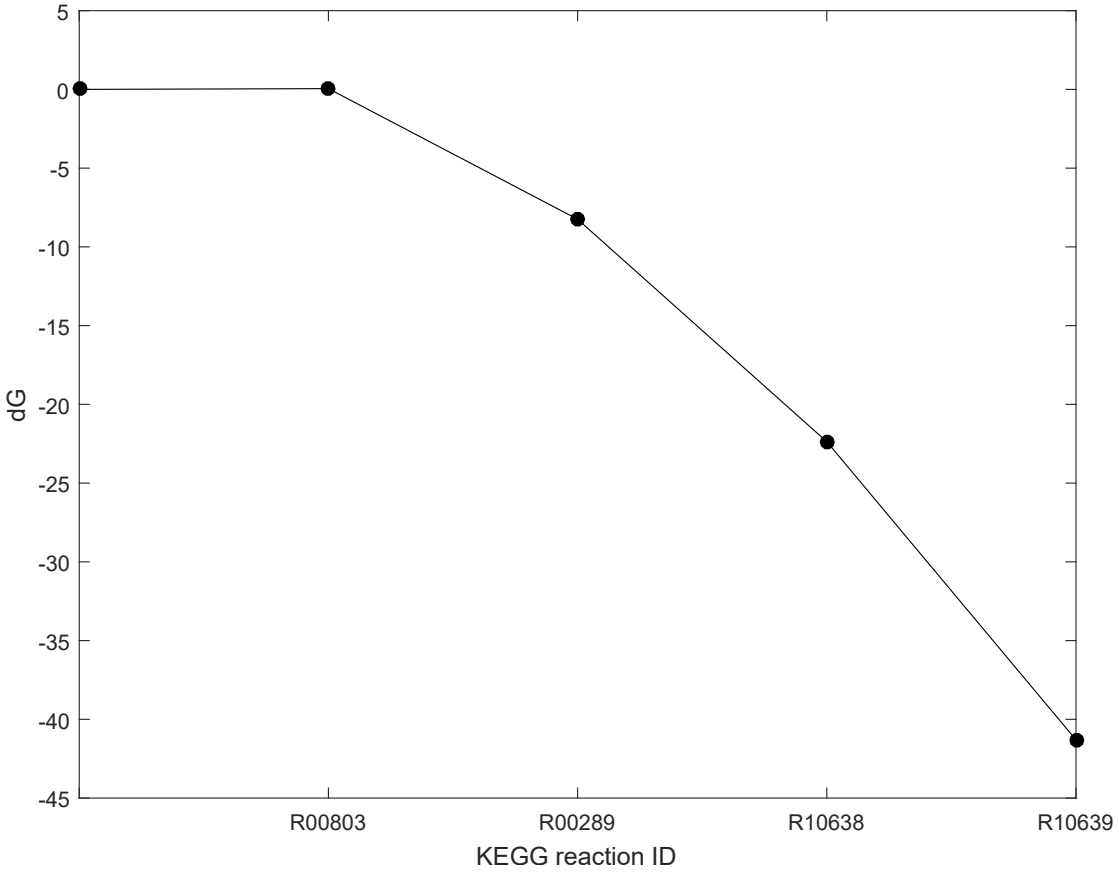


Figure A.1: Thermodynamic profile for the pathway candidate of the synthesis of amygdalin from sucrose.

Overall Balance

Table A.6 shows the overall balance of the pathway candidate for the synthesis of amygdalin from sucrose.

Side Reactions

Table A.6: Overall balance of the pathway candidate to amygdalin starting from sucrose

	R00803	R00289	R10638	R10639	overall
sucrose	-1	0	0	0	-1
orthophosphate	-1	0	0	0	-1
D-fructose	1	0	0	0	1
D-glucose 1-phosphate	1	-1	0	0	0
UTP	0	-1	0	0	-1
diphosphate	0	1	0	0	1
UDP-glucose	0	1	-1	-1	-1
mandelonitrile	0	0	-1	0	-1
prunasin	0	0	0	-1	-1
UDP	0	0	1	1	2
amygdalin	0	0	0	1	1

```
R00291 : UDP-glucose 4-epimerase

  substrates:
    1 C00029  UDP-glucose

  products:
    1 C00052  UDP-alpha-D-galactose

R00959 : alpha-D-Glucose 1-phosphate 1,6-phosphomutase

  substrates:
    1 C00103  D-Glucose 1-phosphate

  products:
    1 C00668  alpha-D-Glucose 6-phosphate

R08639 : alpha-D-glucose 1,6-phosphomutase

  substrates:
    1 C00103  D-Glucose 1-phosphate

  products:
    1 C00092  D-Glucose 6-phosphate

-R00878 : -alpha-D-Glucose aldose-ketose-isomerase

  substrates:
    1 C00095  D-Fructose

  products:
    1 C00267  alpha-D-Glucose
```

Pathway Candidate: Starting From α-d-glucose 6-phosphate

Table A.7: Ranking of pathway candidate to amygdalin from α-D-glucose 6-phosphate

criterion	value
number of active reactions:	4
starts with basic:	False
reactions w/o dG:	0
sum (dG + |dG|):	2.070859e+00
dG:	-3.202206e+01
number of heterologous enzymes:	2
number of cofactors:	1
number of side reactions:	5

```
-R00959 : -alpha-D-Glucose 1-phosphate 1,6-phosphomutase  4.881509e-02

  substrates:
    1 C00668  alpha-D-Glucose 6-phosphate

  products:
    1 C00103  D-Glucose 1-phosphate

R00289 : UTP:alpha-D-glucose-1-phosphate uridylyltransferase  -1.412458e+01

  substrates:
    1 C00075  UTP
    1 C00103  D-Glucose 1-phosphate

  products:
    1 C00013  Diphosphate
    1 C00029  UDP-glucose

R10638 : UDP-D-glucose:(R)-mandelonitrile beta-D-glucosyltransferase  -1.893291e
+01

  substrates:
    1 C00561  Mandelonitrile
    1 C00029  UDP-glucose

  products:
    1 C00844  Prunasin
    1 C00015  UDP

R10639 :  9.866146e-01
```

```
substrates:
  1 C00844  Prunasin
  1 C00029  UDP-glucose

products:
  1 C08325  Amygdalin
  1 C00015  UDP
```

Thermodynamic Profile

Figure A.2 shows the thermodynamic profile for the pathway candidate for the synthesis of amygdalin from α-D-glucose 6-phosphate.

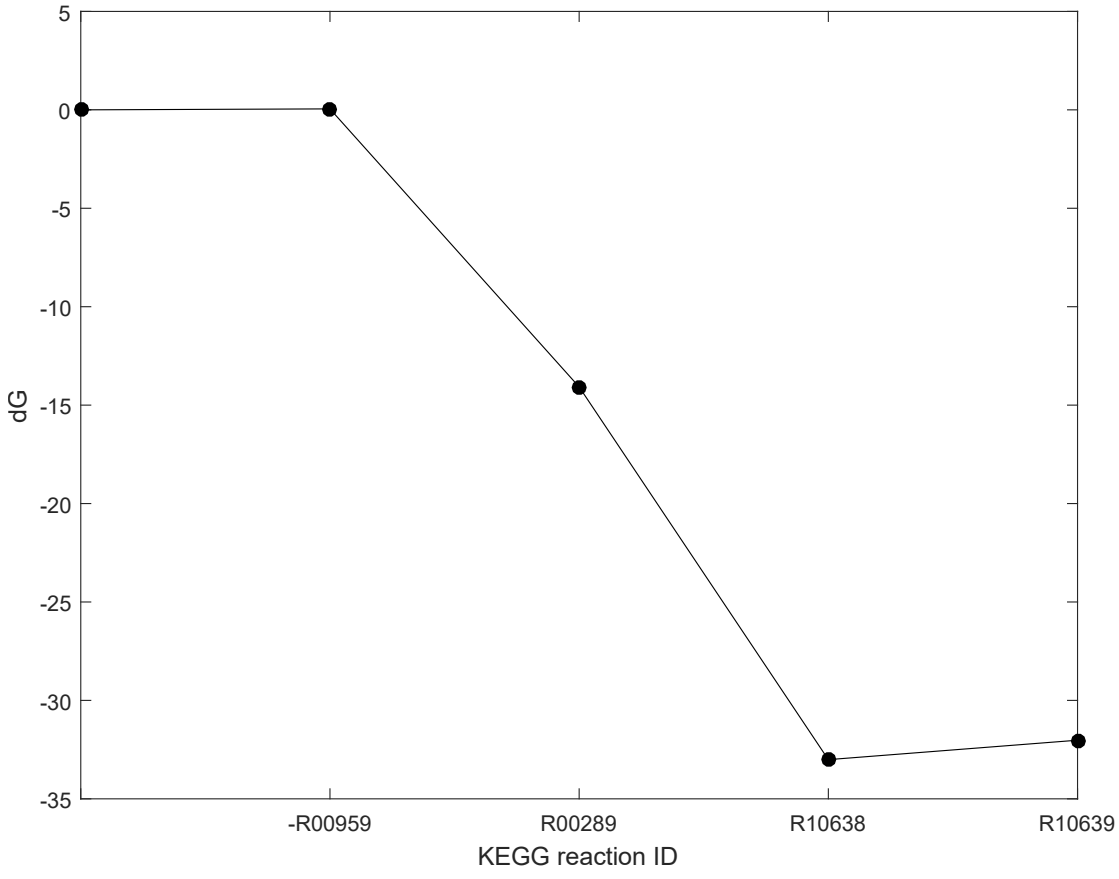


Figure A.2: Thermodynamic profile for the pathway candidate of the synthesis of amygdalin from α-D-glucose 6-phosphate.

Overall Balance

Table A.8 shows the overall balance of the pathway candidate for the synthesis of amygdalin from α-D-glucose 6-phosphate.

Table A.8: Overall balance of the pathway candidate to amygdalin starting from α-D-glucose 6-phosphate

	R00289	R10638	R10639	-R00959	overall
α-D-glucose 6-phosphate	0	0	0	-1	-1
D-glucose 1-phosphate	-1	0	0	1	0
UTP	-1	0	0	0	-1
UDP	0	1	1	0	2
diphosphate	1	0	0	0	0
UDP-glucose	1	-1	-1	0	-1
mandelonitrile	0	-1	0	0	-1
prunasin	0	1	-1	0	0
amygdalin	0	0	1	0	1

Side Reactions

```
R00291 : UDP-glucose 4-epimerase

  substrates:
    1 C00029  UDP-glucose

  products:
    1 C00052  UDP-alpha-D-galactose

R00959 : alpha-D-Glucose 1-phosphate 1,6-phosphomutase

  substrates:
    1 C00103  D-Glucose 1-phosphate

  products:
    1 C00668  alpha-D-Glucose 6-phosphate

R02737 : UDPglucose:D-glucose-6-phosphate 1-alpha-D-glucosyltransferase

  substrates:
    1 C00029  UDP-glucose
    1 C00668  alpha-D-Glucose 6-phosphate

  products:
    1 C00015  UDP
```

```
    1 C00689  alpha,alpha'-Trehalose 6-phosphate

R02740 : alpha-D-Glucose 6-phosphate ketol-isomerase

  substrates:
    1 C00668  alpha-D-Glucose 6-phosphate

  products:
    1 C05345  beta-D-Fructose 6-phosphate

R08639 : alpha-D-glucose 1,6-phosphomutase

  substrates:
    1 C00103  D-Glucose 1-phosphate

  products:
    1 C00092  D-Glucose 6-phosphate
```

Pyrrolysine

Pathway Candidate

Table A.9: Ranking of pathway candidate to pyrrolysine.

criterion	value
number of active reactions:	4
starts with basic:	True
reactions w/o dG:	4
sum (dG + |dG|):	0
dG:	0
number of heterologous enzymes:	4
number of cofactors:	2
number of side reactions:	9

```
R10010 : L-lysine carboxy-aminomethylmutase 0

  substrates:
    1 C00047  L-Lysine

  products:
    1 C20277  (2R,3R)-3-Methylornithine

R10011 :  0
```

```
  substrates:
    1 C00047  L-Lysine
    1 C20277  (2R,3R)-3-Methylornithine
    1 C00002  ATP

  products:
    1 C20278  (2R,3R)-3-Methylornithinyl-N6-lysine
    1 C07305  Products of ATP breakdown

R10012 :  0

  substrates:
    1 C20278  (2R,3R)-3-Methylornithinyl-N6-lysine
    1 C00003  NAD+
    1 C00001  H2O

  products:
    1 C20279  (2R,3R)-3-Methylglutamyl-5-semialdehyde-N6-lysine
    1 C00014  Ammonia
    1 C00004  NADH

R10013 :  0

  substrates:
    1 C20279  (2R,3R)-3-Methylglutamyl-5-semialdehyde-N6-lysine

  products:
    1 C16138  L-Pyrrolysine
    1 C00001  H2O
```

Overall Balance

Table A.10 shows the overall balance of the pathway candidate for the synthesis of pyrrolysine from L-Lysine.

Side Reactions

```
R00086 : ATP phosphohydrolase

  substrates:
    1 C00002  ATP
    1 C00001  H2O

  products:
    1 C00008  ADP
```

Table A.10: Overall balance of the pathway candidate to pyrrolysine.

	R10010	R10011	R10012	R10013	overall
L-lysine	-1	-1	0	0	-2
(2R,3R)-3-methylornithine	1	-1	0	0	0
ATP	0	-1	0	0	-1
(2R,3R)-3-methylornithinyl-N6-lysine	0	1	-1	0	0
products of ATP breakdown	0	1	0	0	1
NAD^+	0	0	-1	0	-1
H_2O	0	0	-1	0	1
(2R,3R)-3-methylglutamyl-5-semialdehyde-N6-lysine	0	0	1	-1	0
NADH	0	0	1	0	1
L-pyrrolysine	0	0	0	1	1

```
    1 C00009  Orthophosphate

R00087 : ATP diphosphohydrolase (diphosphate-forming)

  substrates:
    1 C00002  ATP
    1 C00001  H2O

  products:
    1 C00020  AMP
    1 C00013  Diphosphate

R00089 : ATP diphosphate-lyase (cyclizing

  substrates:
    1 C00002  ATP

  products:
    1 C00575  3',5'-Cyclic AMP
    1 C00013  Diphosphate

R00103 : NAD+ phosphohydrolase

  substrates:
    1 C00003  NAD+
    1 C00001  H2O

  products:
    1 C00020  AMP
    1 C00455  Nicotinamide D-ribonucleotide
```

```
R00104 : ATP:NAD+ 2'-phosphotransferase

  substrates:
    1 C00002  ATP
    1 C00003  NAD+

  products:
    1 C00008  ADP
    1 C00006  NADP+

R00143 : ammonia:NAD+ oxidoreductase

  substrates:
    1 C00014  Ammonia
    1 C00003  NAD+
    1 C00001  H2O

  products:
    1 C00192  Hydroxylamine
    1 C00004  NADH
    1 C00080  H+

R00462 : L-lysine carboxy-lyase (cadaverine-forming)

  substrates:
    1 C00047  L-Lysine

  products:
    1 C01672  Cadaverine
    1 C00011  CO2

R00787 : ammonia:NAD+ oxidoreductase

  substrates:
    1 C00014  Ammonia
    2 C00001  NAD+
    3 C00003  H2O

  products:
    1 C00088  Nitrite
    3 C00004  NADH
    3 C00080  H+

R11104 : NADH phosphohydrolase

  substrates:
    1 C00004  NADH
```

```
    1 C00001  H2O

  products:
    1 C00020  AMP
    1 C21113  NMNH
```

(S)-2-phenyloxirane

Pathway Candidate

Table A.11: Ranking of pathway candidate to (S)-2-phenyloxirane

criterion	value
number of active reactions:	4
starts with basic:	False
reactions w/o dG:	1
sum (dG + |dG|):	2.081496e+01
dG:	-1.728943e+01
number of heterologous enzymes:	4
number of cofactors:	3
number of side reactions:	11

```
R02506 : cinnamaldehyde:NADP+ oxidoreductase (CoA-cinnamoylating) 6.126122e+00

  substrates:
    1 C00903  Cinnamaldehyde
    1 C00010  CoA
    1 C00006  NADP+

  products:
    1 C00540  Cinnamoyl-CoA
    1 C00005  NADPH
    1 C00080  H+

-R02255 : -trans-Cinnamate:CoA ligase (AMP-forming) 0

  substrates:
    1 C00020  AMP
    1 C00013  Diphosphate
    1 C00540  Cinnamoyl-CoA

  products:
    1 C00002  ATP
    1 C00423  trans-Cinnamate
```

```
    1 C00010  CoA

R11070 : trans-cinnamate carboxy-lyase  -2.769690e+01

  substrates:
    1 C00423  trans-Cinnamate

  products:
    1 C07083  Styrene
    1 C00011  CO2

R05488 : styrene,FADH2:oxygen oxidoreductase  4.281356e+00

  substrates:
    1 C07083  Styrene
    1 C01352  FADH2
    1 C00007  Oxygen

  products:
    1 C20782  (S)-2-Phenyloxirane
    1 C00016  FAD
    1 C00001  H2O
```

Thermodynamic Profile

Overall Balance

Table A.12 shows the overall balance of the pathway candidate to (S)-2-phenyloxirane.

Side Reactions

```
R00004 : diphosphate phosphohydrolase

  substrates:
    1 C00013  Diphosphate
    1 C00001  H2O

  products:
    2 C00009  Orthophosphate

R00086 : ATP phosphohydrolase

  substrates:
    1 C00002  ATP
    1 C00001  H2O
```

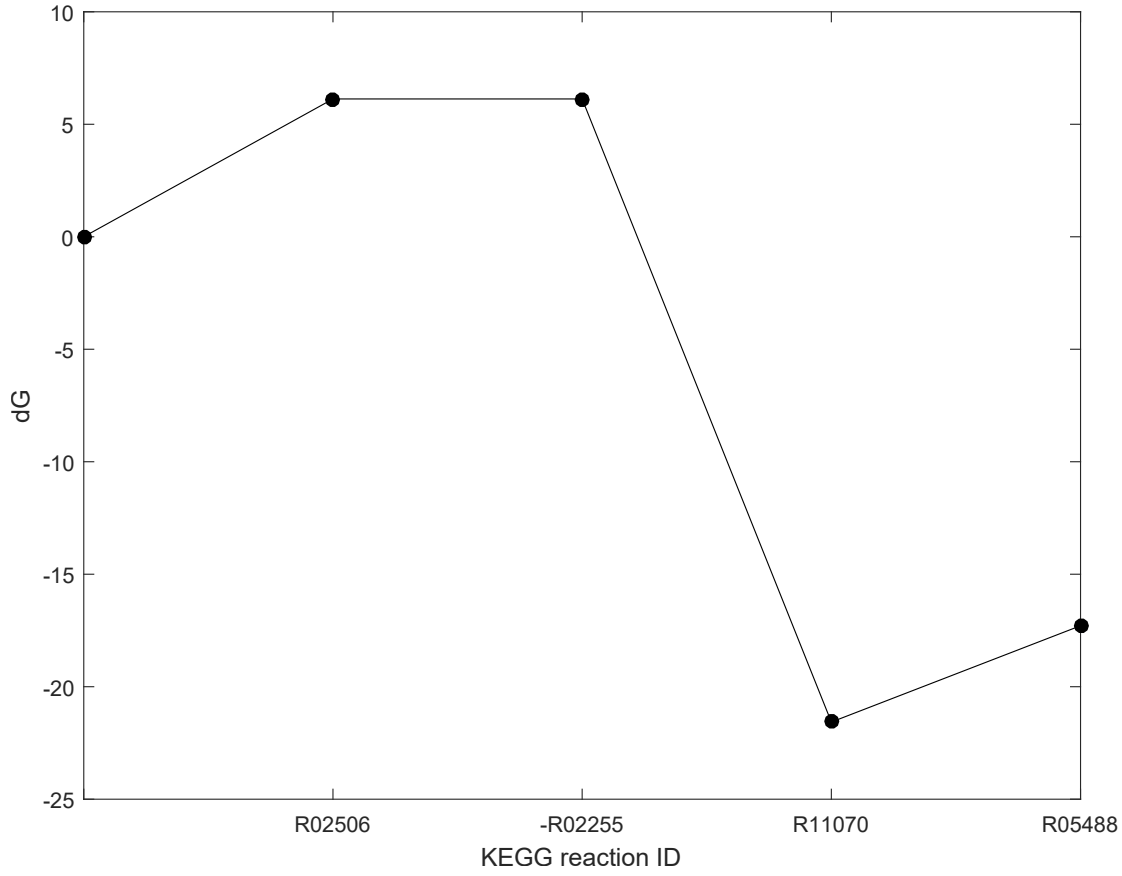


Figure A.3: Thermodynamic profile for a pathway candidate of the synthesis of (S)-2-phenyloxirane from cinnamaldehyde.

```
  products:
    1 C00008  ADP
    1 C00009  Orthophosphate

R00087 : ATP diphosphohydrolase (diphosphate-forming)

  substrates:
    1 C00002  ATP
    1 C00001  H2O

  products:
    1 C00020  AMP
    1 C00013  Diphosphate

R00089 : ATP diphosphate-lyase (cyclizing
```

Table A.12: Overall balance of the pathway candidate to (S)-2-phenyloxirane.

	R02506	-R02255	R11070	R05488	overall
cinnamaldehyde	-1	0	0	0	-1
CoA	-1	1	0	0	0
$NADP^+$	-1	0	0	0	-1
cinnamoyl-CoA	1	-1	0	0	0
NADPH	1	0	0	0	1
H^+	1	0	0	0	1
AMP	0	-1	0	0	-1
diphosphate	0	-1	0	0	-1
ATP	0	1	0	0	1
trans-cinnamate	0	1	-1	0	0
styrene	0	0	1	-1	0
CO_2	0	0	0	1	1
$FADH_2$	0	0	0	-1	-1
oxygen	0	0	0	-1	-1
(S)-2-phenyloxirane	0	0	0	1	1
FAD	0	0	0	1	1
H_2O	0	0	0	1	1

```
  substrates:
    1 C00002  ATP

  products:
    1 C00575  3',5'-Cyclic AMP
    1 C00013  Diphosphate

R00127 : ATP:AMP phosphotransferase

  substrates:
    1 C00002  ATP
    1 C00020  AMP

  products:
    2 C00008  ADP

R00182 : AMP phosphoribohydrolase

  substrates:
    1 C00020  AMP
    1 C00001  H2O

  products:
    1 C00147  Adenine
```

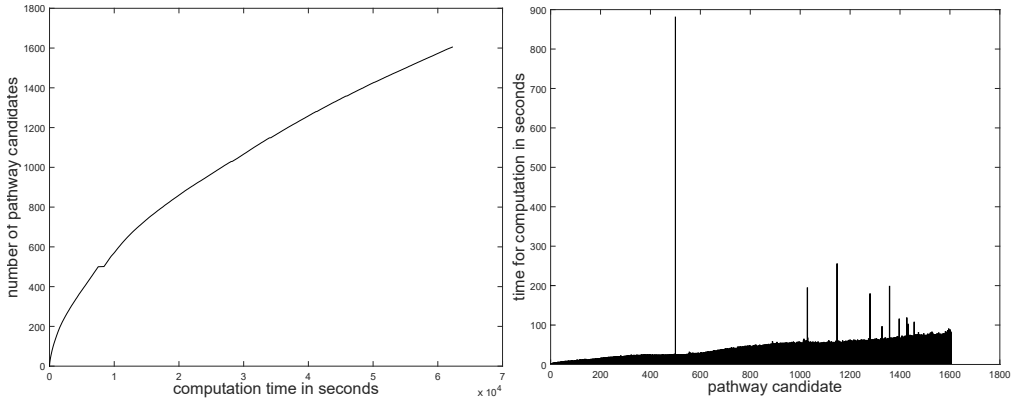
```
    1 C00117  D-Ribose 5-phosphate

R00183 : adenosine 5'-monophosphate phosphohydrolase

  substrates:
    1 C00020  AMP
    1 C00001  H2O

  products:
    1 C00212  Adenosine
    1 C00009  Orthophosphate

R00190 : AMP:diphosphate phospho-D-ribosyltransferase

  substrates:
    1 C00020  AMP
    1 C00013  Diphosphate

  products:
    1 C00147  Adenine
    1 C00119  5-Phospho-alpha-D-ribose 1-diphosphate

-R00132 : -carbonate hydro-lyase (carbon-dioxide-forming)

  substrates:
    1 C00011  CO2
    1 C00001  H2O

  products:
    1 C01353  Carbonic acid

-R00161 : -ATP:FMN adenylyltransferase

  substrates:
    1 C00013  Diphosphate
    1 C00016  FAD

  products:
    1 C00002  ATP
    1 C00061  FMN

-R10092 : -carbonate hydro-lyase (carbon-dioxide-forming)

  substrates:
    1 C00011  CO2
    1 C00001  H2O

  products:
```

```
1 C00288  HC03-
1 C00080  H+
```

A.2 Computation Times

Geranyl Pyrophosphate

We generated 1645 pathway candidates. Figure A.4(a) shows the number of pathway candidates over time. The computation time for the individual pathway candidates is shown in Figure A.4(b).



(a) Number of pathway candidates over time for geranyl pyrophosphate.

(b) Computation time for individual pathway candidates for geranyl pyrophosphate.

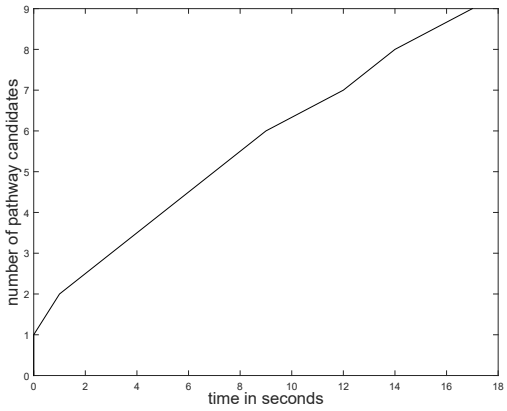
Figure A.4: Geranyl pyrophosphate.

Amygdalin

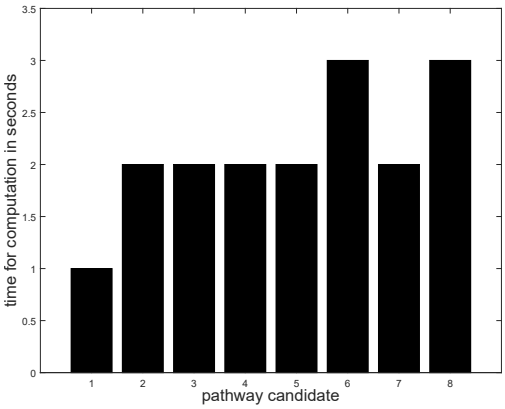
We generated 100 pathway candidates. Figure A.5(a) shows the number of pathway candidates over time. The computation time for the individual pathway candidates is shown in Figure A.5(b).

Pyrrolysine

5 pathway candidates were generated. Figure A.6(a) shows the number of pathway candidates over time. The computation time for the individual pathway candidates is shown in Figure A.6(b).

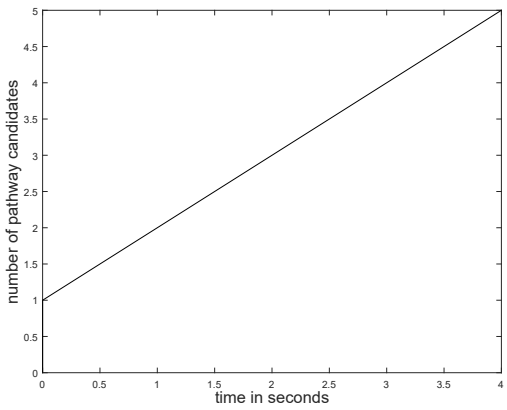


(a) Computation time for individual pathway candidates for amygdalin.

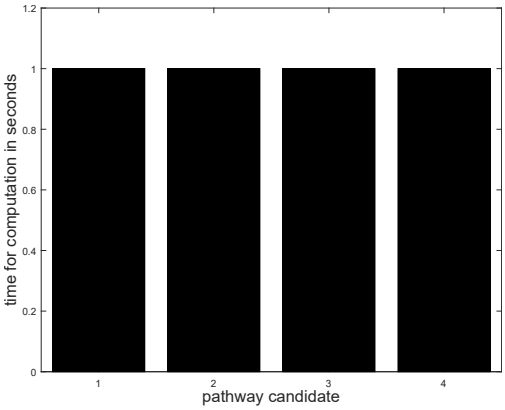


(b) Number of pathway candidates over time for amygdalin.

Figure A.5: Amygdalin.



(a) Number of pathway candidates over time for pyrrolysine.



(b) Computation time for individual pathway candidates for pyrrolysine.

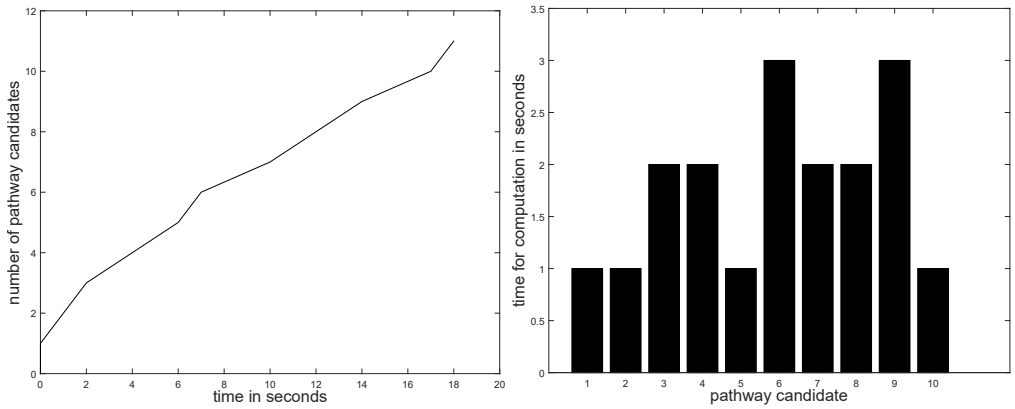
Figure A.6: Pyrrolysine.

(S)-2-phenyloxirane

11 pathway candidates were generated. Figure A.7(a) shows the number of pathway candidates over time. The computation time for the individual pathway candidates is shown in Figure A.7(b).

Remarks

We showed the number of pathway candidates over time empirically for the example pathway searches we presented in Section 4.3 Results. For a typical pathway search, the computation



(a) Number of pathway candidates over time for (S)-2-phenyloxirane.

(b) Computation time for individual pathway candidates for (S)-2-phenyloxirane.

Figure A.7: (S)-2-phenyloxirane.

time was in the range of minutes for the first 100 pathway candidates. For all examples, the trend shows that the solver takes more time with each solution, due to the fact that for each additional solution, the number of constraint and variables in the MILP grows.

B Appendix In-depth characterization of genome-scale network reconstructions for the *in vitro* synthesis in cell-free systems

The following chapter is based on the the supplementary material of the research article (SCHUH et al., 2019) (Chapter 5 Network Reconstructions for Cell-Free Systems).

B.1 MILP

The MILP presented below identifies pathway candidates for a given target T from a given list of possible starting metabolites.

$$\sum_{i=1}^{|M|} u_{iP} = 1 \tag{B.1}$$

$$\sum_{j=1}^{|M|} u_{Tj} = 0 \tag{B.2}$$

$$\sum_{i=1}^{|M|} u_{il} \leq \sum_{j=1}^{|M|} u_{lj} \quad l \in S; \quad l \neq T \tag{B.3}$$

$$\sum_{i=1}^{|M|} u_{il} = 0 \quad l \in B; \quad l \neq T \tag{B.4}$$

$$\sum_{i=1}^{|M|} u_{ik} = \sum_{j=1}^{|M|} u_{kj} \quad k \in M \setminus S; \quad k \neq T \tag{B.5}$$

$$\sum_{i=1}^{|M|} u_{ik} \leq 1, \quad k = 1, \ldots, |M| \tag{B.6}$$

Constraints (B.1) to (B.6) ensure that a solution contains a connected simple path from a start node of the set of designated start nodes to the given end note T.

$$\sum_{r=1}^{|R|} S_{mr} v_r \geq 0, \quad \forall m \in E, m \notin E_m \tag{B.7}$$

$$\sum_{r=1}^{|R|} S_{Tr} v_r \geq 1, \tag{B.8}$$

$$z_r \leq v_r, \qquad r = 1, \ldots, R \tag{B.9}$$

$$\text{and} \quad v_r \leq Max \cdot z_r, \qquad r = 1, \ldots, R \tag{B.10}$$

$$z_\lambda + z_\mu \leq 1 \tag{B.11}$$

$\forall(\lambda,\mu) \in B = \{(\lambda,\mu) | \lambda$ and μ are reverse$\}$

$$\sum_{r=1}^{|R|} d_{ijr} \cdot z_r \geq u_{ij} \quad i = 1, \ldots, |M|; j = 1, \ldots, |M|; i \neq j \tag{B.12}$$

The constraint formulated in equation (B.13) prevents the use of a a reaction in the pathway that consumes the target T. The set R_T is the set of reactions that consume the target and thus have a negative stoichiometric coefficient.

$$\sum_{r \in R_T} v_r S_r \geq 0, \quad R_T = \{r | S_{Tr} < 0\} \tag{B.13}$$

Constraints (B.7) to (B.13) define a valid flux distribution for the pathway ensuring that the found path is feasible.

The objective function is given in equation (B.14).

$$Minimize \quad \sum_{i=1}^{|M|} \sum_{j=1, j\neq i}^{|M|} u_{ij} + \frac{1}{|R|+1} \sum_{i=1}^{|R|} z_i \tag{B.14}$$

The MILP consisting of equations (B.1) to (B.14) provides a pathway candidate given by a sequence of arcs (i.e. the values of u_{ij}) and the active reactions (the values of z_r). The objective function guarantees a connected and cycle-free linear path with a minimal number of supplying reactions.

The remaining constraints are used to find further pathway candidates.

$$\sum_{i}^{|M|} \sum_{j}^{|M|} U_{ij}^{k'} \cdot s_{k'} \leq \sum_{i}^{|M|} \sum_{j}^{|M|} U_{ij}^{k'} u_{ij} \tag{B.15}$$

$$\sum_{i}^{|M|} \sum_{j}^{|M|} (1 - U_{ij}^{k'}) u_{ij} + s_{k'} |M|^2 \leq |M|^2 \tag{B.16}$$

$$\sum_{i}^{|M|} \sum_{j}^{|M|} U_{ij}^{k'} u_{ij} - \sum_{i}^{|M|} u_{i\alpha^{k'}} - s_{k'} \leq \sum_{i}^{|M|} \sum_{j}^{|M|} U_{ij}^{k'} - 1 \tag{B.17}$$

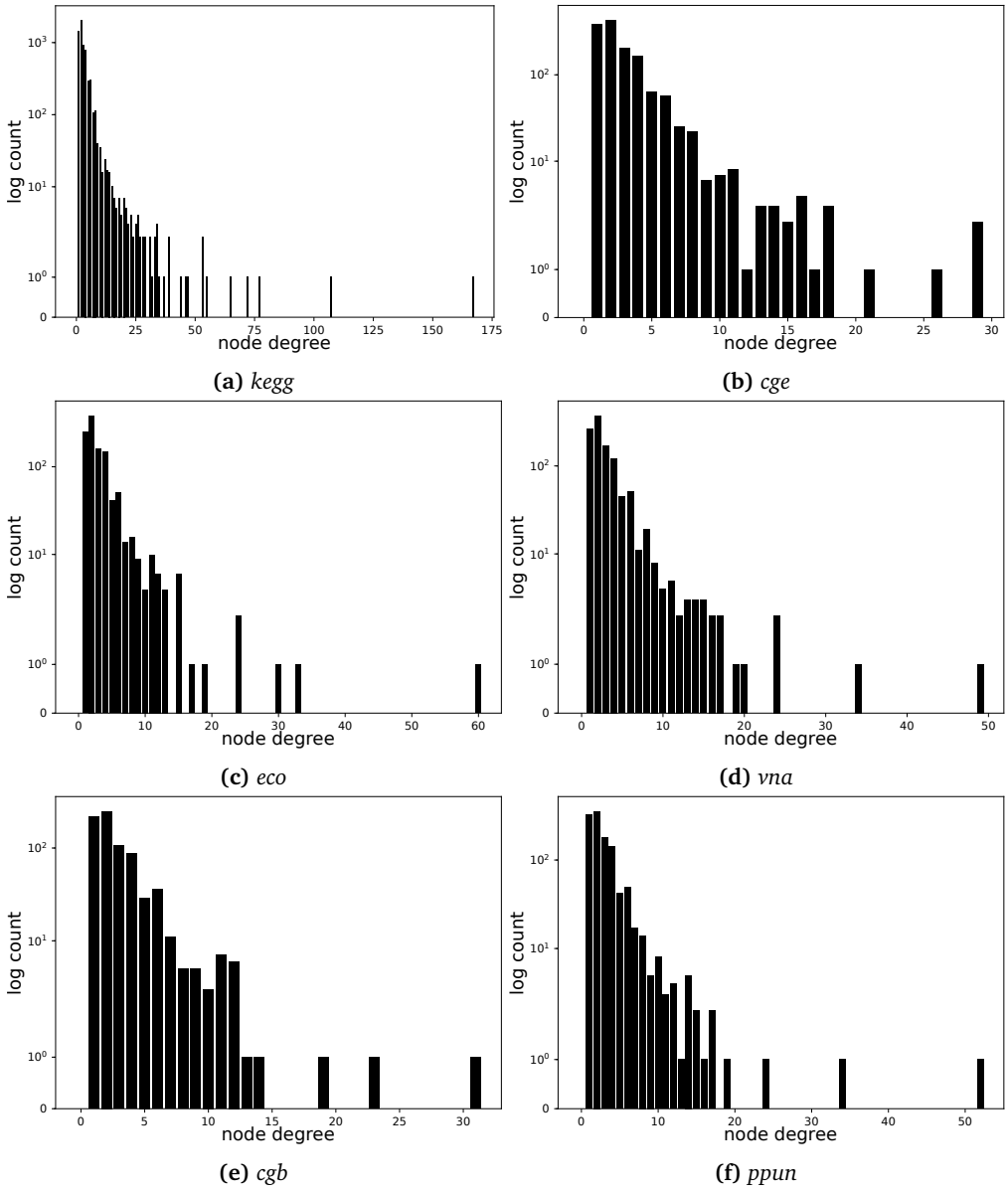
$$\sum_{i}^{|R|} Z_i^{'l} z_i + s_k |R| \leq m_l - 1 + |R| \tag{B.18}$$

An in-depth discussion of each constraint of the MILP can be found in Section 4.2 Mathematical Model.

B.2 Arc Graph Properties

Node Degree Distributions

Figure B.1 shows the node degree distributions of the arc graphs of the different organism network reconstructions.

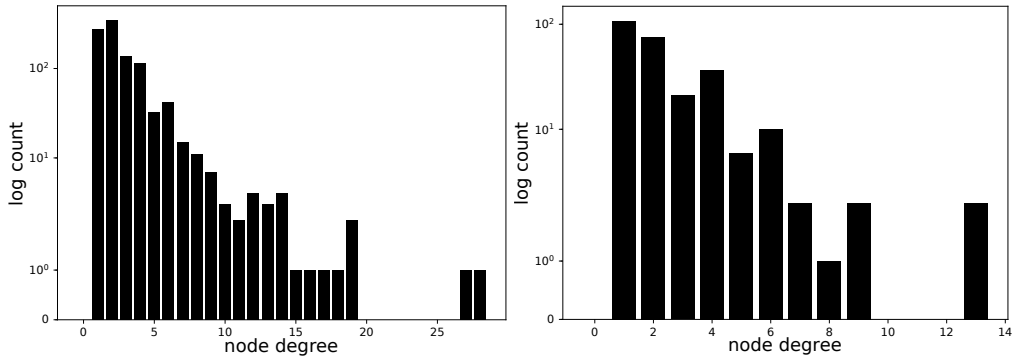


(a) *kegg* **(b)** *cge*

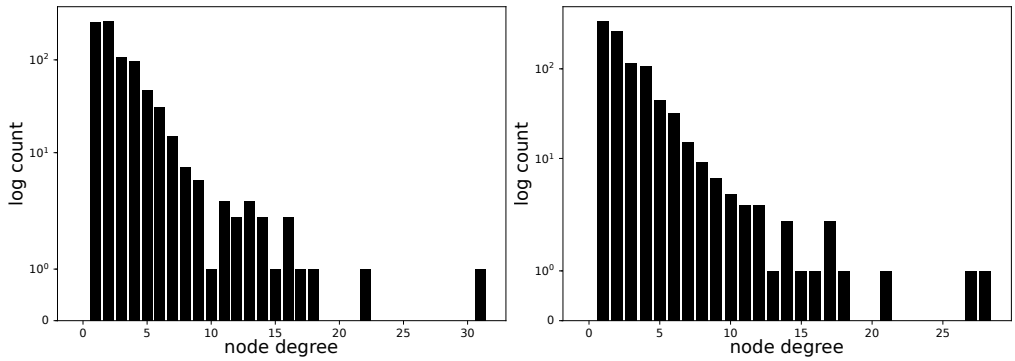
(c) *eco* **(d)** *vna*

(e) *cgb* **(f)** *ppun*

Figure B.1: Total node degree distributions of the different organism networks.



(g) *mxa* **(h)** *mpe*



(i) *spo* **(j)** *sce*

Figure B.1: Total node degree distributions of the different organism networks (continued).

Hubs

Tables B.1 to B.10 list the hubs with the top 5 occurrences of each network.

Table B.1: Hubs in *kegg*

count	KEGG ID	name
167	C00022	pyruvate
107	C00448	*trans*,*trans*-farnesyl diphosphate
77	C00033	acetate
72	C00031	D-glucose
65	C00025	L-glutamate

Model Properties

Table B.11 lists the sizes of the arc graphs together with the average node degrees (sum of ingoing and outgoing arcs) and the standard deviation σ.

Table B.2: Hubs in *cge*

count	KEGG ID	name
29	C00022	pyruvate
29	C00025	L-glutamate
26	C00024	acetyl-CoA
21	C00037	glycine
18	C05345	β-D-fructose 6-phosphate
18	C00020	AMP
18	C00065	L-serine

Table B.3: Hubs in *eco*

count	KEGG ID	name
60	C00022	pyruvate
33	C00024	acetyl-CoA
30	C00111	glycerone phosphate
24	C00025	L-glutamate
24	C00118	D-glyceraldehyde 3-phosphate

Table B.4: Hubs in *vna*

count	KEGG ID	name
49	C00022	pyruvate
34	C00024	acetyl-CoA
24	C00025	L-glutamate
24	C00111	glycerone phosphate
20	C00118	D-glyceraldehyde 3-phosphate

Table B.5: Hubs in *ppun*

count	KEGG ID	name
52	C00022	pyruvate
34	C00024	acetyl-CoA
24	C00025	L-glutamate
19	C00118	D-glyceraldehyde 3-phosphate
17	C00109	2-oxobutanoate

Table B.6: Hubs in *mxa*

count	KEGG ID	name
28	C00024	acetyl-CoA
27	C00022	pyruvate
19	C00025	L-glutamate
19	C00118	D-glyceraldehyde 3-phosphate
18	C00037	L-serine

Table B.7: Hubs in *sce*

count	KEGG ID	name
28	C00022	pyruvate
27	C00024	acetyl-CoA
21	C00025	L-glutamate
18	C05345	β-D-fructose 6-phosphate
17	C00118	D-glyceraldehyde 3-phosphate
17	C00065	L-Serine

Table B.8: Hubs in *spo*

count	KEGG ID	name
31	C00022	pyruvate
22	C00025	L-glutamate
18	C05345	β-D-fructose 6-phosphate
17	C00118	D-glyceraldehyde 3-phosphate
16	C00020	AMP
16	C00026	2-oxoglutarate

Table B.9: Hubs in *cgb*

count	KEGG ID	name
31	C00022	pyruvate
23	C00025	L-glutamate
19	C00118	D-glyceraldehyde 3-phosphate
14	C00024	acetyl-CoA
13	C00111	glycerone phosphate

Table B.10: Hubs in *mpe*

count	KEGG ID	name
13	C00111	glycerone phosphate
13	C00118	D-glyceraldehyde 3-phosphate
9	C00085	D-fructose 6-phosphate
9	C05345	β-D-fructose 6-phosphate
8	C03794	N6-(1,2-dicarboxyethyl)-AMP
7	C05378	β-D-fructose 1,6-bisphosphate
7	C00147	adenine

Table B.11: Number of metabolites, arcs and average node degrees (sum of ingoing and outgoing arcs) and their standard deviation σ of each model.

model	metabolites with arcs	arcs	average node degree	σ
kegg	6246	10291	3.3	4.45
cge	1395	2056	2.95	2.64
eco	1106	1736	3.14	3.28
vna	1093	1663	3.04	3.04
ppun	1154	1683	2.92	2.9
mxa	997	1424	2.86	2.58
sce	954	1303	2.73	2.55
spo	840	1187	2.83	2.54
cgb	768	1093	2.85	2.51
mpe	263	312	2.37	1.83

One can observe that the average node degrees are similar for all models, regardless of the arc graph size. The higher average node degree in *kegg* could be caused by the fact that this model consists of all reactions fulfilling the criteria discussed in Section 5.2, whereas in the organism models the reactions are additionally selected by organism annotation, which may introduce a bias.

Arc Graph Creation

As explained in Section 5.2 Model Building, the basis of the arc graph is the set of reactant pairs from KEGG RCLASS for each reaction without those containing cofactors or inorganic metabolites. In the following, we will discuss the reasoning behind that choice. To do so, we built four different arc graphs for each model, listed in Table B.12. Arc graph 1

Table B.12: Setup for the four different arc graphs in discussion. In the second column, the '+' marks that the arcs in the respective arc graph are derived from the KEGG RCLASS entries of the reactions in the model, the '-' denotes that the cross product of all substrates and products of the reactions is used. In the third column, a '+' means that arcs containing cofactors and inorganic compounds are included, a '-' means that such arcs are excluded.

arc graph	RCLASS	cofactors/inorganics	number arcs	number vertices
1	+	-	10291	6246
2	+	+	10697	6329
3	-	-	15118	6366
4	-	+	25231	6467

is built as described in Section 5.2 Model Building and comprises arcs derived from the RCLASS reaction pairs (i.e. reaction pairs based on chemical structure patterns), without arcs

containing cofactors or inorganic metabolites. This arc graph was applied in our path-finding calculations. This setup allows synthesis of the predicted compounds in the network, as demonstrated by the results presented and discussed in Section 5.3 Reachability Analysis. In arc graph 2, we additionally included arcs involving cofactors and inorganic metabolites. As KEGG RCLASS predominantly contains relevant arcs, arcs involving cofactors and inorganic metabolites are not normally included in the first place. For this reason the number of arcs does not significantly increase from arc graph 1 to 2. In arc graph 3 all possible substrate-reaction pairs are used, but arcs involving cofactors are not, whereas arc graph 4 consists of all substrate-reaction pairs. For each of those arc graphs we investigated the top 10 hubs (based on occurrence), which are listed for the pan-organism network *kegg*, exemplarily in Tables B.13 to B.16. These tables show that those arc graphs which incorporate arcs containing containing cofactors and inorganic metabolites (arc graphs 2 and 4) also have such metabolites in their top 10 hubs. In the arcs derived from all substrate-reaction pairs, 8 out of 10 hubs are cofactors or inorganic metabolites. For the arcs derived from KEGG RCLASS, only one of the ten hubs is a cofactor. It is thus reasonable to exclude cofactors and inorganic metabolites from arcs, as already stated in Section 5.2 Model Building. Comparing the arc graphs 1 and 3 (Tables B.13 and B.15), one can observe that arc graph 3, which is based on all substrate-reaction pairs, contains mono- and diphosphates such as AMP or ADP. These metabolites are not present in the top 10 hubs of arc graph 1, which is based on KEGG RCLASS. Let us have a look at the example reaction given in equation 1.

$$\mathrm{S} + \mathrm{ATP} \rightleftharpoons \mathrm{P} + \mathrm{ADP} + \mathrm{AMP} \qquad \{1\}$$

Here, all substrate-reaction pairs are S-P, S-ADP, S-AMP, ATP-P, ATP-ADP, ATP-AMP. In RCLASS most likely only the S-P arc will appear. Arc graph 3 uses S-P, S-ADP, S-AMP. In most cases the arcs S-ADP and S-AMP are not meaningful connections, however. The additional arcs in arc graph 3 predominantly consist of this kind of nonsensical arcs. It is thus reasonable to derive the arcs from KEGG RCLASS and exclude arcs with cofactors and inorganic metabolites.

Target Statistics

Table B.17 lists the numbers used for the analysis of the target search in the different organism models in Figure 5.6.

Table B.13: Top 10 hubs for arc graph 1 of Table B.12 based on KEGG RCLASS without arcs containing cofactors and inorganic metabolites. The first column gives the number of occurrences, the second column the KEGG COMPOUND id and the last column the name of the metabolite.

count	KEGG id	name
167	C00022	pyruvate
107	C00448	*trans*,*trans*-farnesyl diphosphate
77	C00033	acetate
72	C00031	D-glucose
65	C00025	L-glutamate
55	C00067	formaldehyde
53	C00024	acetyl-CoA
53	C00058	formate
47	C00084	acetaldehyde
46	C00037	glycine

Table B.14: Top 10 hubs for arc graph 2 of Table B.12 based on KEGG RCLASS including arcs containing cofactors or inorganic metabolites. The first column gives the number of occurrences, the second column the KEGG COMPOUND id and the last column the name of the metabolite. KEGG ids in bold are cofactors or inorganic metabolites.

count	KEGG ID	name
167	C00022	pyruvate
147	**C00010**	CoA
107	C00448	*trans*,*trans*-farnesyl diphosphate
77	C00033	acetate
72	C00031	D-glucose
67	C00025	L-glutamate
56	C00058	formate
55	C00024	acetyl-CoA
55	C00067	formaldehyde
49	C00037	glycine

Table B.15: Top 10 hubs for arc graph 3 of Table B.12 based on all substrate-reaction pairs without arcs containing cofactors and inorganic metabolites. The first column gives the number of occurrences, the second column the KEGG COMPOUND id and the last column the name of the metabolite.

count	KEGG ID	name
2445	C00022	pyruvate
519	C00008	ADP
489	C00021	S-Adenosyl-L-homocysteine
316	C00026	2-oxoglutarate
289	C00015	UDP
242	C00025	L-glutamate
234	C00024	acetyl-CoA
219	C00020	AMP
205	C00029	UDP-glucose
157	C00042	succinate

Table B.16: Top 10 hubs for arc graph 4 of Table B.12 based on all substrate-reaction pairs including arcs containing cofactors or inorganic metabolites. The first column gives the number of occurrences, the second column the KEGG COMPOUND id and the last column the name of the metabolite. KEGG ids in bold are cofactors or inorganic metabolites.

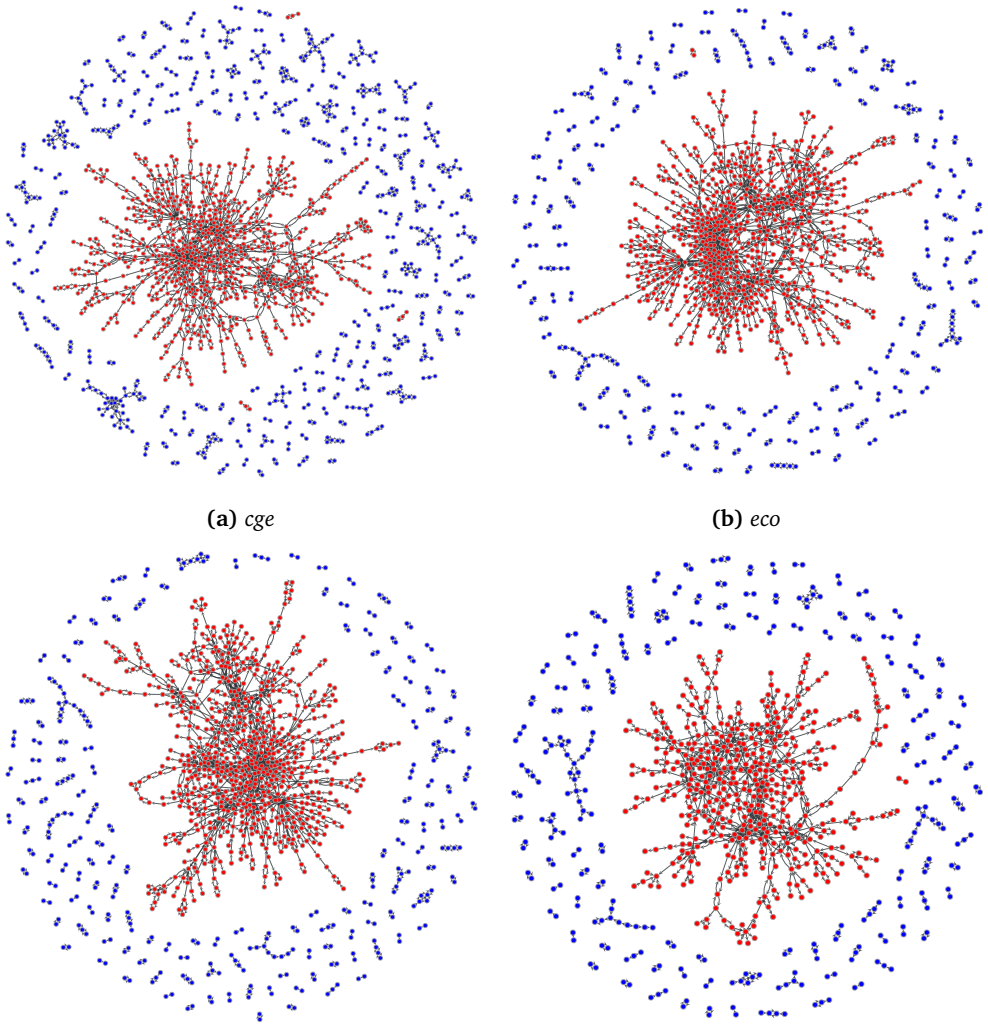
count	KEGG ID	name
1892	**C00001**	H_2O
1315	**C00080**	H^+
1062	**C00007**	oxygen
699	**C00005**	NADPH
620	**C00006**	$NADP^+$
617	**C00004**	NADH
577	**C00003**	NAD^+
545	C00008	ADP
500	C00021	S-Adenosyl-L-homocysteine
449	**C00002**	ATP

Table B.17: Raw data for plot in Figure 5.6. *found and not predicted*: targets for which a target candidate has been found by our method, but that have not been predicted as feasible; *found and predicted*: targets for which a target candidate has been found by our method; *not found (connectivity)*: targets for which a candidate has not been found due to the absence of an arc path from any start metabolite to the target; *not found (feasibility)*: targets for which a candidate has not been found due to the lack of a feasible reaction that produces the target; *not found (feasibility)*: targets for which a candidate has not been found due to other reasons that are discussed in Section 5.3 Reachability Analysis.

	kegg	*cge*	*eco*	*vna*	*ppun*	*mxa*	*sce*	*spo*	*cgb*	*mpe*
found and not predicted	99	9	15	1	0	2	4	0	11	0
found and predicted	2294	317	335	315	287	250	213	186	185	48
not found (connectivity - satellite components)	1342	534	275	291	291	313	354	301	242	87
not found (connectivity - components with start metabolites)	1082	157	184	177	231	144	91	72	95	27
not found (feasibility)	593	95	53	68	87	52	37	63	50	14
not found (other)	31	16	16	13	6	16	14	15	15	8

Arc Graph Maps

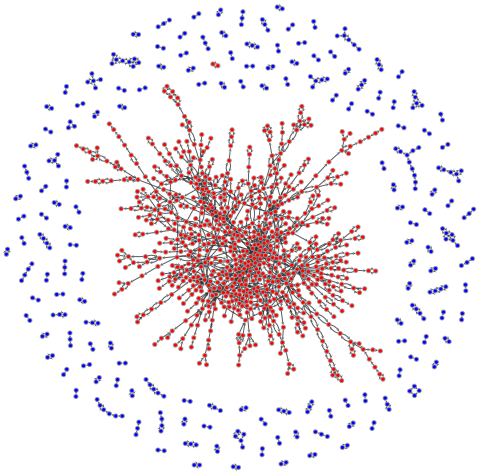
Figure B.2 shows the arc graphs of the different organism networks. The components colored in red are components containing potential start metabolites. Components in blue are components without start metabolites.



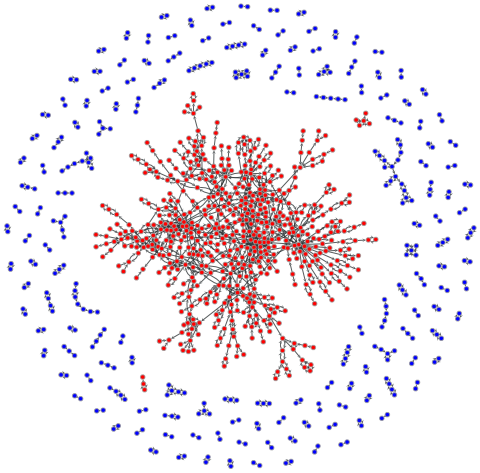
(a) *cge* **(b)** *eco*

(c) *vna* **(d)** *cgb*

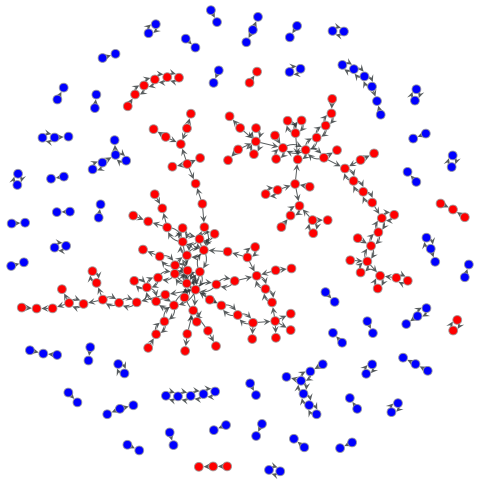
Figure B.2: Arc graphs of the different organism networks. Red: components containing potential start metabolite; blue: satellite components without start metabolites.



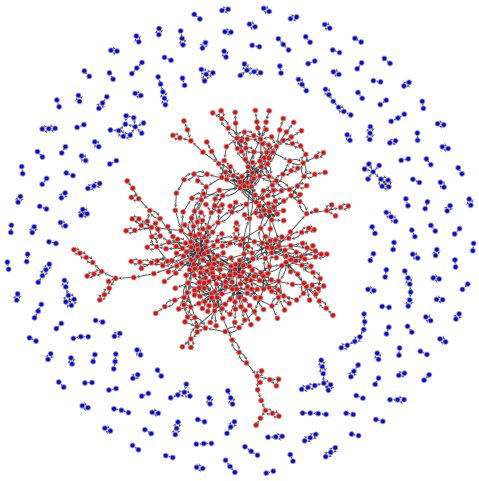
(e) *ppun*



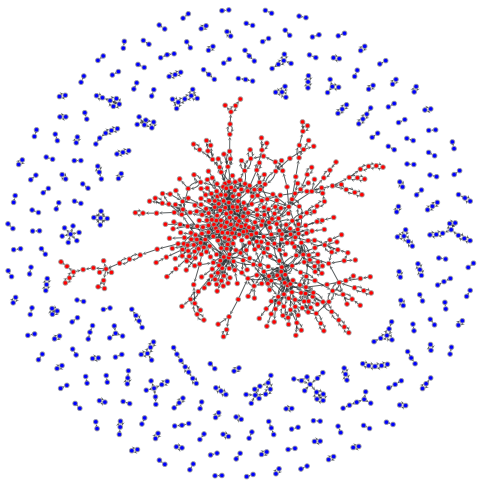
(f) *mxa*



(g) *mpe*



(h) *spo*

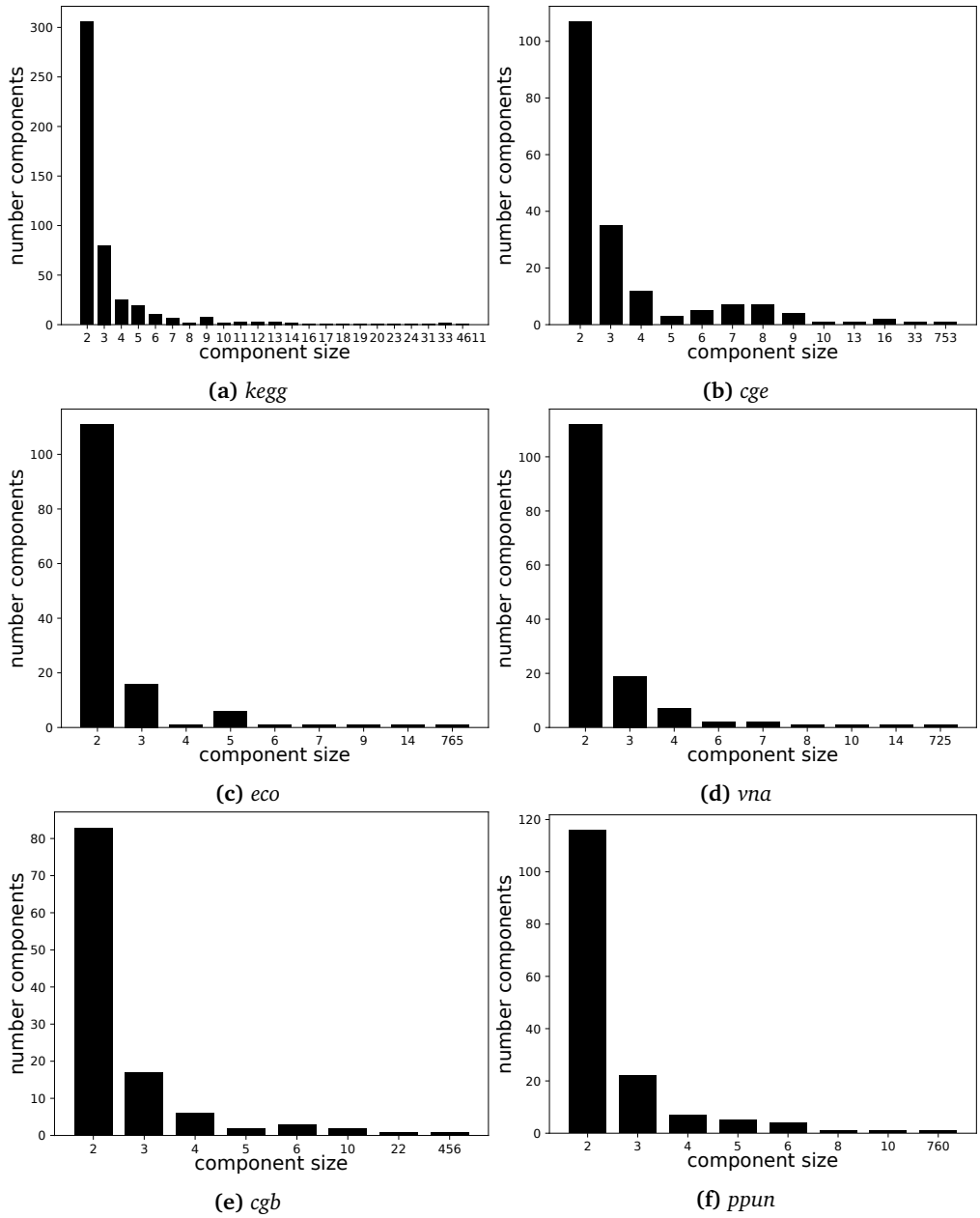


(i) *sce*

Figure B.2: Arc graphs of the different organism networks (continued). Red: components containing potential start metabolite; blue: satellite components without start metabolites.

Arc Graph Component Histograms

Figure B.3 shows the arc graph component histograms of the different organism networks.

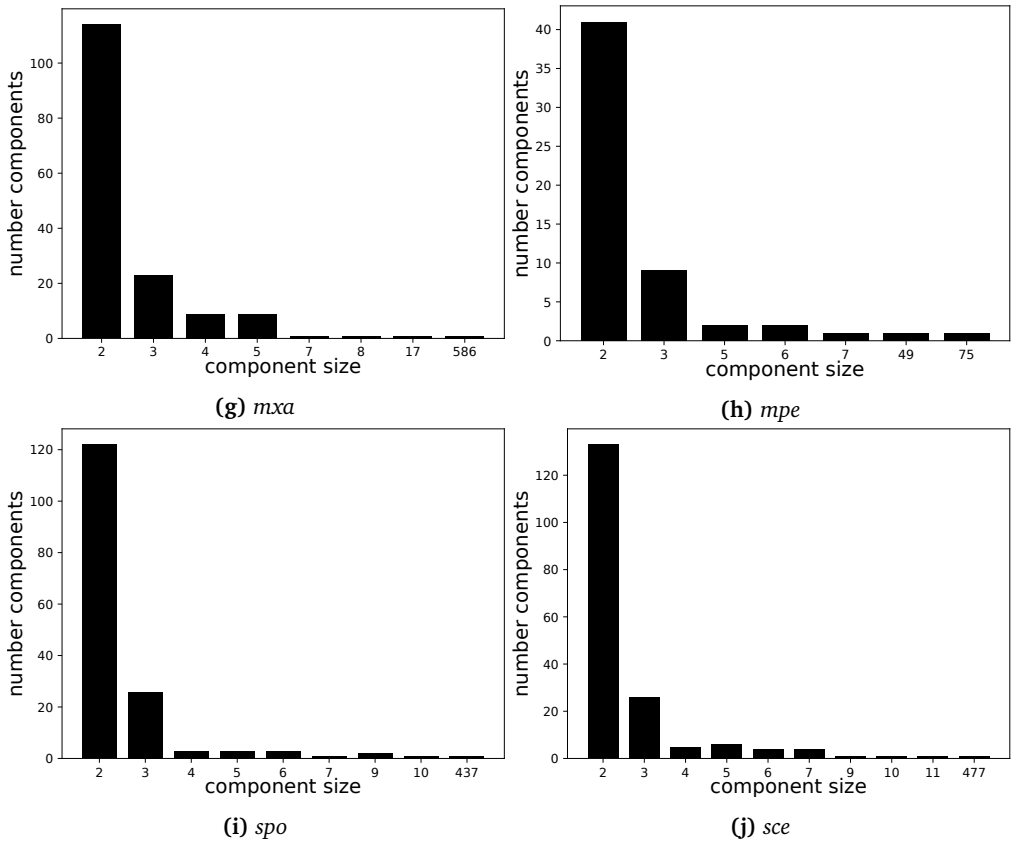


(a) *kegg* **(b)** *cge*

(c) *eco* **(d)** *vna*

(e) *cgb* **(f)** *ppun*

Figure B.3: Arc graph component histograms of the different organism networks.



(g) *mxa* **(h)** *mpe*

(i) *spo* **(j)** *sce*

Figure B.3: Arc graph component histograms of the different organism networks (continued).

B.3 Components of the *kegg* Model

The components identified for the *kegg* model are listed in Table B.18. Component 1 contained 4612 compounds/metabolites representing the central metabolic network. In total there are 481 components with 6246 compounds/metabolites.

We investigated the identified components not directly connected to the main component 1 that contains 4612 compounds in the range from 5 to 33 compounds (Supplementary Information 3 of (SCHUH et al., 2019)). These are in total 69 components representing 682 compounds/metabolites.

The members of the larger remaining components were of the families of
polyketides including macrolides (21/23, 28/19, 74/13, 63/10, 94/9, 115/9, 217/9, 152/6, 156/6, 112/5) - 10
flavonoids (15/33, 263/9, 210/7, 247/5) - 4

Table B.18: Components in model kegg

Number of components	Size of component (number of compounds)
306	2
80	3
25	4
19	5
10	6
7	7
2	8
8	9
2	10
3	11
3	12
3	13
2	14
1	16
1	17
1	18
1	19
1	20
1	23
1	24
1	31
2	33
1	4612

terpenoids (6/33, 7/31, 185/13, 38/7, 47/7, 245/6, 27/5, 223/5, 270/5) - 9
compounds with gonane type nucleus, also known as perhydrocyclopenta[a]phenanthrene (e.g. steroids, cholic acid derivatives): (49/18, 5/17, 66/16, 73/7, 168/6, 31/5, 64/5) - 7
xenobiotic compounds, e.g. drugs (32/16, 108/14, 95/12, 109/12, 11/11, 18/11, 36/11, 70/9, 147/9, 206/9, 116/8, 23/7, 114/7, 194/5) - 14
other degradation pathways (30/12, 305/6) - 2
carbohydrate derived (4/24, 134/13, 68/9, 61/8, 180/6, 260/6, 83/5, 101/5, 149/5, 250/5) - 10
fatty acids and lipids related (25/20, 44/5, 83/5, 101/5, 234/5, 318/5, 318/5) - 7
porphyrins (22/10, 62/6) - 2
amino acid derived (127/7, 69/6, 253/6, 9/5, 188/5) - 5
nucleotide derived (137/5) - 1

B.4 Target Examples

UDP-glucose

```
-R00806 : -R00806 UDP-glucose:D-fructose 2-alpha-D-glucosyltransferase  dG:
0.02295

  substrates:
    1 C00089  Sucrose
    1 C00015  UDP

  products:
    1 C00029  UDP-glucose
    1 C00095  D-Fructose

R00159 : R00159 UTP phosphohydrolase  dG: -41.17

  substrates:
    1 C00001  H2O
    1 C00075  UTP

  products:
    1 C00009  Orthophosphate
    1 C00015  UDP
```

5-methyl-5,6,7,8-tetrahydromethanopterin

```
R04456 : R04456 5,10-methylenetetrahydromethanopterin:coenzyme-F420
oxidoreductase  dG: 0

substrates:
  1 C00876  Coenzyme F420
  1 C04377  5,10-Methylenetetrahydromethanopterin
  1 C00080  H+

products:
  1 C01080  Reduced coenzyme F420
  1 C04330  5,10-Methenyltetrahydromethanopterin

R09099 : R09099 5,10-methylenetetrahydromethanopterin:glycine
hydroxymethyltransferase  dG: -9.909

substrates:
  1 C00065  L-Serine
  1 C01217  5,6,7,8-Tetrahydromethanopterin
```

```
products:
  1 C00037  Glycine
  1 C04377  5,10-Methylenetetrahydromethanopterin
  1 C00001  H2O

R04464 : R04464 5,10-Methylenetetrahydromethanopterin:coenzyme-F420
oxidoreductase  dG: -10.2

substrates:
  1 C01080  Reduced coenzyme F420
  1 C04377  5,10-Methylenetetrahydromethanopterin

products:
  1 C00876  Coenzyme F420
  1 C04488  5-Methyl-5,6,7,8-tetrahydromethanopterin
```

Biotin

Table B.19 lists reactions in the model that produce biotin but do not belong to the feasible reactions in the network.

Table B.19: Reactions producing biotin that are not feasible in the network.

KEGG id	reaction equation
R10127	biotin sulfoxide + NAD(P)H + H^+ <=> biotin + $NADP^+$ + H_2O
-R01075	AMP + diphosphate + biotinyl-CoA <=> ATP + biotin + CoA
R01076	biotin amide + H_2O <=> biotin + ammonia
R01077	biocytin + H_2O <=> biotin + L-lysine

5'-methylthioadenosine

In the following, we list reactions in our model that produce 5'-methylthioadenosine and belong to the feasible reactions in the network.

```
R10881
S-Adenosyl-L-methionine + Nocardicin G <=> 5'-Methylthioadenosine + Isonocardicin
 C

no arc to target

R00180
S-Adenosyl-L-methionine <=> 5'-Methylthioadenosine + Homoserine lactone

no arc (cofactors)
```

no arc to target

R11089
S-Adenosylmethioninamine + Norspermine <=> 5'-Methylthioadenosine +
Caldopentamine + H+

no pathway to substrate

-R01402
Adenine + S-Methyl-5-thio-D-ribose 1-phosphate <=> 5'-Methylthioadenosine +
Orthophosphate

no pathway to substrate

R08359
S-Adenosylmethioninamine + Cadaverine <=> 5'-Methylthioadenosine +
Aminopropylcadaverine

no pathway to substrate

R10338
S-Adenosylmethioninamine + Agmatine <=> 5'-Methylthioadenosine + N1-(3-
Aminopropyl)agmatine

no pathway to substrate

R01920
S-Adenosylmethioninamine + Putrescine <=> 5'-Methylthioadenosine + Spermidine

no pathway to substrate

R00175
S-Adenosyl-L-methionine + H2O <=> 5'-Methylthioadenosine + L-Homoserine

no arc (cofactors)
no arc to target

R11088
S-Adenosylmethioninamine + Norspermidine <=> 5'-Methylthioadenosine + Norspermine
+ H+

no pathway to substrate

R11154
2 S-Adenosylmethioninamine + Spermidine <=> 2 5'-Methylthioadenosine + N4-Bis(
aminopropyl)spermidine

```
no pathway to substrate

R00179
S-Adenosyl-L-methionine <=> 1-Aminocyclopropane-1-carboxylate + 5'-
Methylthioadenosine

no arc (cofactors)
no arc to target

R11159
S-Adenosylmethioninamine + Spermidine <=> 5'-Methylthioadenosine + N4-
Aminopropylspermidine

no pathway to substrate

R03726
S-Adenosyl-L-methionine + N6-(Delta2-Isopentenyl)-adenine <=> 5'-
Methylthioadenosine + Discadenine

no arc to target

R03271
S-Adenosylmethioninamine + 1,3-Diaminopropane <=> 5'-Methylthioadenosine +
Norspermidine + H+

no pathway to substrate

R09531
S-Adenosylmethioninamine + Spermidine <=> 5'-Methylthioadenosine + Thermospermine
 + H+

no pathway to substrate

R03072
S-Adenosyl-L-methionine + Nocardicin E <=> 5'-Methylthioadenosine + Isonocardicin
 A

no arc to target

R02869
S-Adenosylmethioninamine + Spermidine <=> 5'-Methylthioadenosine + Spermine

no pathway to substrate
```

C Appendix: Synthesis Paths for UDP-glucose

C.1 Pathway Candidates

```
sucrose -> UDP-glucose

-R00806 : -R00806 UDP-glucose:D-fructose 2-alpha-D-glucosyltransferase  dG: 1.06

  substrates:
    1 C00089  Sucrose
    1 C00015  UDP

  products:
    1 C00029  UDP-glucose
    1 C00095  D-Fructose

R00158 : R00158 ATP:UMP phosphotransferase  dG: 4.863

  substrates:
    1 C00105  UMP
    1 C00002  ATP

  products:
    1 C00015  UDP
    1 C00008  ADP
```

Listing 1: Two-step pathway candidate to UDP-glucose from sucrose.

```
sucrose -> D-glucose 1-phosphate -> UDP-glucose

R00803 : R00803 sucrose:phosphate alpha-D-glucosyltransferase dG: -8.474

  substrates:
    1 C00009  Orthophosphate
    1 C00089  Sucrose

  products:
    1 C00103  D-Glucose 1-phosphate
    1 C00095  D-Fructose
```

```
R00568 : R00568 CTP aminohydrolase  dG: -42.62

  substrates:
    1 C00063  CTP
    1 C00001  H2O

  products:
    1 C00014  Ammonia
    1 C00075  UTP

R00289 : R00289 UTP:alpha-D-glucose-1-phosphate uridylyltransferase dG: 0.929

  substrates:
    1 C00103  D-Glucose 1-phosphate
    1 C00075  UTP

  products:
    1 C00013  Diphosphate
    1 C00029  UDP-glucose
```

Listing 2: Pathway candidate to UDP-glucose from sucrose.

```
R00156 : R00156 ATP:UDP phosphotransferase  dG: -2.68

  substrates:
    1 C00002  ATP
    1 C00015  UDP

  products:
    1 C00008  ADP
    1 C00075  UTP

R00158 : R00158 ATP:UMP phosphotransferase  dG: -4.863

  substrates:
    1 C00105  UMP
    1 C00002  ATP

  products:
    1 C00015  UDP
    1 C00008  ADP

R00289 : R00289 UTP:alpha-D-glucose-1-phosphate uridylyltransferase dG: 0.929

  substrates:
    1 C00103  D-Glucose 1-phosphate
    1 C00075  UTP
```

```
  products:
    1 C00013  Diphosphate
    1 C00029  UDP-glucose

-R08639 : -R08639 alpha-D-glucose 1,6-phosphomutase dG: 7.392

  substrates:
    1 C00092  D-Glucose 6-phosphate

  products:
    1 C00103  D-Glucose 1-phosphate
```

Listing 3: Two-step pathway candidate to UDP-glucose from D-glucose 6-phosphate.

D Software

The software developed for this study is divided into two packages. The first package comprises the tools for model building and model analysis. It is for the most part implemented in Python 2.7; except the thermodynamics part, which is implemented in Python 3.6. The implementation makes use of the external packages `eQuilibrator` (FLAMHOLZ et al., 2012), `graph-tool` (PEIXOTO, 2014) and `scipy` (JONES et al., 2001). The second package is the implementation of the path-finding algorithm and the pathway ranking and analysis, written in MATLAB R2019a using IBM ILOG CPLEX as MILP solver.

D.1 Model Building

KEGG Parser

Parsers for extracting the data needed for model building from the raw KEGG entries from KEGG COMPOUND, REACTION and ENZYME databases were implemented. All parsers take a raw text entry as input and create an instance of the respective class containing the relevant data for the model. The parsers can be found in the following scripts in `MECATPy_KEGG`: `compound.py`, `reaction.py`, `enzyme.py`, `rclass.py`, `organism.py` and `pathway.py`. Parsing the raw data can be invoked by calling the script `KEGGreader.py`, which only needs the respective filenames for the raw data and automatically creates Python pickles and MATLAB dictionaries containing the parsed data.

Model

The network reconstruction models consist of the files listed in Table D.1, which are generated by the script `buildHostModels.py`.

Table D.1: Model files. *model* denotes the chosen model name (typically the KEGG organism code).

filename	description
cofactors*model*.txt	textfile containing the internal ids of the dedicated cofactor and inorganic compounds
cofactors_names*model*.txt	textfile containing the KEGG ids, names and internal ids of the dedicated cofactor and inorganic compounds
compounds*model*.txt	textfile containing the KEGG ids of all compounds contained in the model
compounds_names*model*.txt	textfile containing the KEGG ids, names and internal ids of all compounds contained in the model
start_compounds*model*.txt	textfile containing the KEGG ids of the designated start compounds
start_compounds_names*model*.txt	textfile containing the KEGG ids and names of the designated start compounds
terminal_compounds*model*.txt	textfile containing the KEGG ids of the designated basic compounds
terminal_compounds_names*model*.txt	textfile containing the KEGG ids and names of the designated basic compounds
targets_organism.txt	textfile containing the KEGG ids of the potential target metabolites
targets_organism_names.txt	textfile containing the KEGG ids and names of the potential target metabolites
reactions*model*.txt	textfile containing the KEGG ids of all reactions contained in the model. A '-' in front of the id denotes that the reaction is reversed.
reactions_names*model*.txt	textfile containing the KEGG ids and names of all reactions contained in the model
reversible_reactions*model*.txt	textfile containing the internal ids of the reversible reactions (first column: reaction in KEGG direction, second column: reaction in reversed direction)
reversible_reactions_names*model*.txt	textfile containing the KEGG ids of the reversible reactions
non_enzymatic_reactions*model*.txt	textfile containing the KEGG ids of the non-enzymatic reactions contained in the model. A '-' in front of the id denotes that the reaction is reversed.
non_enzymatic_reactions_names*model*.txt	textfile containing the names of the non-enzymatic reactions contained in the model

Table D.1: Model files. *model* denotes the chosen model name (typically the KEGG organism code) (continued).

filename	description
thermodynamics.mat	MATLAB map containing the thermodynamic data of all model reactions
thermodynamics*model*.map	MATLAB struct containing the thermodynamic data of all model reactions
RxR_kegg*model*.txt	textfile containing the arcs of the model given as pairs of internal compound ids and the internal id of the respective reaction
RxR_kegg_names*model*.txt	textfile containing the arcs of the model given as pairs of KEGG compound ids and the KEGG id of the respective reaction
arcs.mat	MATLAB matrix containing the arcs of the model given as pairs of internal compound ids
S*model*.mat	MATLAB matrix with the stoichiometrix matrix. Row and column indices correspond to the internal compound and reaction indices
*model*_model.pickle	Python pickle file containing the `Model` class instance of the respective model
feasible_reactions.txt	textfile with the KEGG ids and equations of the feasible reactions (with respect to the given metabolite pool)

D.2 Path-Finding

Input File

All necessary parameters for the path-finding tool can be given in a special input file, described in Table D.2.

Table D.2: Parameters for the input file for the path-finding tool.

parameter	type	comment
DATA_PATH	string	path to the model data
GENERIC_METABOLITES	text file	contains generic metabolites
COFACTORS	text file	contains internal ids of cofactors
START_METABOLITES	text file	contains internal ids of start metabolites
BASIS_METABOLITES	text file	contains internal ids of basis metabolites
COMPOUNDS	text file	contains KEGG compound ids
REACTIONS	text file	contains KEGG reaction ids
COMPOUND_NAMES	text file	contains compound names
REACTION_NAMES	text file	contains reaction names
ARCS	text file	contains reaction pairs
STOICHIOMETRIC_MATRIX	mat	MATLAB formatted data containing the stoichiometric matrix
REVERSIBLE_REACTIONS	text file	contains reversible reactions
KEGG_DATA_PATH	string	path to the KEGG data
REACTION_MAP	mat	MATLAB formatted data containing reaction data
COMPOUND_MAP	mat	MATLAB formatted data containing compound data
ENZYME_MAP	mat	MATLAB formatted data containing enzyme data
PATHWAY_MAP	mat	MATLAB formatted data containing pathway data
NON_ENZYMATIC_REACTIONS	text file	contains KEGG reaction ids of non-enzymatic reactions
THERMODYNAMICS_MAP	mat	MATLAB formatted data containing thermodynamics data
SBML_MODEL	xml	SBML file of model
REVERSIBILITIES_MAP	mat	MATLAB formatted data containing reaction reversibility information
HOST	string	KEGG organism code

MILP Generation and Path-Finding

The basic script for the MILP generation and path-finding is `fsr.m`. It takes the filename of the aforementioned input file. The basic MILP is set up with the script `buildBasicProblem.m`, which is called automatically by `fsr.m`. The additional constraints to exclude pathways that have already been found are added successively by the script `findSynthesisRoutesAll.m` (also called automatically). There are two modes for path-finding. The default mode is running the solver until no new pathway candidates can be found. The second one searches for a user defined number of pathway candidates (at most). This can be set up in the script `findSynthesisRoutesAll.m`. All pathway candidates are stored as MATLAB `*.mat` files in the predefined folder under the respective target KEGG id and can be analyzed with the ranking scripts described in the following section.

D.3 Analysis

The analysis encompasses different aspects. The first aspect is the ranking of the pathway candidates for a given target. The second aspect is the analysis of the network reconstructions. The last aspect is the analysis of the results of the pathway searches to all possible potential targets of the network reconstructions.

Ranking

The ranking of pathway candidates for a specific target is done with the MATLAB script `writeTargets.m`, which takes the filename of the configuration input file (Table D.2). The script automatically calls the ranking scripts in the respective order, which can be varied depending on how much weight a specific ranking function should have in the overall ranking. In the default case, the order is as given in Table 4.1. The script also writes the pathway candidates in a user-friendly manner which then can be assessed further.

Model Statistics

The Python script `Statistics\ModelStatistics.py` provides various model statistics. The script writes general metabolite statistics including the number of metabolites, number of external/generic metabolites, the size of the metabolite pool, number of start and basis metabolites, as well as number of cofactors/inorganic metabolites. The reaction statistics encompass the number of reactions in the model, number of unique reactions (without reverse) and number of reversible reactions.

There are analysis methods that explore properties of the arc graph of the network reconstructions. For example, the script writes a list of the network metabolites together with the component to which they belong to (Supplementary Information 3 of (SCHUH et al., 2019)). This information can also be divided into lists containing all metabolites of the main component and the metabolites of the other component. It also generates a histogram. The degree distribution of the arc graphs can also be plotted. The script also performs the reachability screening presented in Section 5.2. It also outputs the potential targets that are reachable by BFS and those that are predicted to be feasible.

It is also possible to output the top n hubs in the network depending on the arc construction method. If arcs with cofactors/inorganics are chosen, the script does also output which of the top n hub metabolites belong to this group. Common hubs, which appear in the top n hubs of each network model are also written.

Result Statistics

A prerequisite for the following scripts is that (i) for a given metabolic network reconstruction the existence of a pathway candidate to each possible target is tested using the path-finding method and (ii) the results have been processed with the MATLAB script `writeTargets.m` to generate the file `targets_overview`.

The main part of the Python script `OrganismResultStatistics.py` is the generation of the raw data and the plot shown in Table B.17 and Figure 5.6 based on the results of the path-finding experiment. For each model, lists with the targets belonging to each category are written. The script also determines which targets are common in all organism network reconstructions (including the pan-organism model) and which ones are unique (excluding the pan-organism network model).

List of Publications and Conference Contributions

Scientific Publications

Blass, L. K., Weyler, C., & Heinzle, E. (2017): Network design and analysis for multi-enzyme biocatalysis. *BMC Bioinformatics* (Aug. 2017), vol. *18*([1]): 366. https://doi.org/10.1186/s12859-017-1773-y

Heinzle, E., Weyler, C., Krauser, S., & Blass, L. K. (2013): Directed multistep biocatalysis using tailored permeabilized cells. A.-P. Zeng (Ed.), *Advances in biochemical engineering/biotechnology* (pp. 185–234). Springer Berlin Heidelberg. https://doi.org/10.1007/10_2013_240

Schuh, L. K., Weyler, C., & Heinzle, E. (2019): In-depth characterization of genome-scale network reconstructions for the in vitro synthesis in cell-free systems. *Biotechnology and Bioengineering* (2019), vol. *117*([4]): 1137–1147. https://doi.org/10.1002/bit.27249

Conference Contributions

Poster

Blass, L. K., Weyler, C., & Heinzle, E. (2017): *Computational network design and analysis* [Metabolic Pathway Analysis, Bozeman, Montana USA]. Metabolic Pathway Analysis, Bozeman, Montana USA. http://www.chbe.montana.edu/biochemenglab/documents/17_MPA_program_complete.pdf

Talk

Blass, L. K. (2012): *Network design and analysis for multi-enzyme catalysis* [UniGR I-Derbi - Workshop Systems Biology, Luxembourg]. UniGR I-Derbi - Workshop Systems Biology, Luxembourg.

Danksagung

Ein besonderer Dank gilt Prof. Elmar Heinzle für die Bereitstellung des interessanten Themas und die kompetente und stets unterstützende wissenschaftliche Betreuung.

Ich danke ebenfalls herzlich meinem wissenschaftlichen Begleiter Prof. Volkhard Helms für seine Bereitschaft, diese Arbeit zu begutachten.

Diese Arbeit wurde teilweise finanziell vom BMBF im Rahmen des MECAT Projektes unterstützt.

Einen herzlichen Dank möchte ich an alle Mitarbeitern der Technischen Biochemie richten, die mich herzlich in die Gruppe aufgenommen haben. Hier danke ich besonders Dr. Christian Weyler, der durch viele Diskussionen und Brainstorm-Sessions zum Entstehen dieser Arbeit beigetragen hat und auch über die Arbeit hinaus zu einem guten Freund geworden ist.

Vielen Dank ebenfalls an Dr. Malina Orsini und Yeda Kaminski für viele schöne Stunden im und vor allem außerhalb des Büros.

Schließlich möchte ich auch ganz besonders meiner Familie danken, die durch ihre Unterstützung in jeder Hinsicht zu dieser Arbeit beigetragen haben.