universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Creating an enzymatic database with graph rewrite rules“

verfasst von / submitted by

Bojana Ristivojcevic, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2021 / Vienna 2021

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 066 862
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Chemie
Betreut von / Supervisor:	ao. Univ.-Prof. Mag. Dr. Christoph Flamm

Danksagung

An dieser Stelle möchte ich mich bei meiner Mutter und meinem Vater bedanken, die mir das Studieren ermöglicht haben.

Ich möchte mich für die Betreuung und Unterstützung ganz herzlich bei meinem Betreuer Prof. Dr. Christoph Flamm bedanken.

Weiters gebührt mein Dank meinem Bruder, meinem Freund und meinen Freunden, die mich während des Studiums immer unterstützt und ertragen haben.

Ebenfalls möchte ich mich bei meinen Arbeitskollegen, vor allem Julia am TBI für eine freundliche Atmosphäre bedanken.

Zusammenfassung

Enzyme spielen eine sehr wichtige Rolle im Stoffwechsel des Organismus. Alle Stoffwechselschritte / -prozesse werden durch Enzyme katalysiert.
Ziel der Masterarbeit war es eine Datenbank, welche in einer Webseite öffentlich zugänglich ist, zu erstellen, um Graph Rewrite Regeln für chemisch, enzymatische Reaktionsgraphen zu speichern und bei Bedarf chemische Reaktionsnetzwerke aufzubauen.
Bei der Untersuchung chemischer Reaktionsnetzwerke liegt das Interesse darin zu erkennen, in welchen Atomen der Edukte während einer biochemischen Reaktion zu welchen Atomen der Produkte übergegangen wurde. Graphendatenbanken sind ein geeignetes Werkzeuge für eine effiziente und bequeme Möglichkeit, chemische Netzwerke zu speichern und zu erkunden. In der gegenwärtigen Situation können Netzwerke nicht durch Datenbanken aufgebaut werden, da in den Datenbanken die Reaktionen lediglich als Bilder und Gesamtreaktionsdaten gespeichert werden, die für generative Netzwerkkonstruktionsansätze nicht geeignet sind. Außerdem werden in Reaktionsdatenbanken normalerweise nur Verbindungen und manchmal auch Transformationen aufgelistet, keine Atom Mappings. Die Atom Mappings einer Reaktion beschreibt für jedes Nichtwasserstoffatom in einer Reaktantenverbindung und das entsprechende Atom in einer Produktverbindung. Wenn die Atom Mappings von Reaktion bestimmt werden, kann man Reaktionszentren identifizieren, welche zur Untersuchung von Reaktionsmechanismen verwendet werden können. Der Reaktionsmechanismus beschreibt detailliert die einzelnen Elementarreaktionen der gesamten chemischen Reaktion. So ist deutlich zu erkennen, welche Eduktatome in welche Produktatome umgewandelt wurden. Somit können die Atome der chemischen Reaktionen innerhalb der Reaktion verfolgt werden. Um diesen generativen Ansatz beizubehalten, muss zunächst die Biochemie als Graph-Rewrite-Regel vollständig zugänglich sein, um dann im nächsten Schritt die Reaktionszentren, jene Teil der Reaktion bei dem sich die Bindungen brechen bzw bilden, zu finden. Eine chemische Graph-Rewrite Regel ist eine aggregierte Abstraktion elementarer quantenmechanischer Schritte der Elektronenrekonfiguration, welche stattfinden müssen, um die strukturelle Änderung der chemischen Reaktion umzusetzen. In diesem Sinne muss der Kontext einer Graph-Rewrite-Regel alle notwendigen physikalisch-chemischen Eigenschaften erfassen, die vorhanden sein müssen, damit die elementaren quantenmechanischen Schritte zur Implementierung des Strukturwandels stattfinden können.
Dies wären die ersten Schritte zur datengesteuerten Erzeugung von Reaktionsnetzwerken. Ziel dieser Datenbank ist es auch von Personen verwendet zu werden, die wenig oder keine chemische Erfahrung haben. Eine nützliche Anwendung dieses Ansatzes ist die Syntheseplanung von enzymkatalysierten Pfaden. Aufgrund des Netzwerks sind Synthesewege leichter verständlich und es ist auch klar, welche Transformation durch welches Enzym katalysiert werden kann. Der automatisierte Aufbau von Atomumverteilungsnetzwerken, der durch einen vorgegebenen Satz von Enzymen induziert wird und für die Interperetation von Isoptopenmarkierungsexperimenten im Rahmen der Analyse des Stoffwechselflusses erforderlich ist, ist ein weiterer zukünftiger Anwendungsbereich der Datenbank. Ein weiterer Vorteil wäre, dass die Graph-Rewrite Regeln in der erstellten Datenbank als minimale und maximale Version gespeichert sind und auch immer darauf zugegriffen werden kann. Die maximale Graph-Rewrite

Regel enthält alle Atome mit den Atom Mappings der gesamten Reaktion. Die minimale Graph-Rewrite Regel enthält lediglich diejenigen Atome, bei denen die Bindungen direkt gebildet oder gebrochen werden, also das Reaktionszentrum. So kann der Benutzer von der erstellten Datenbank die Graph-Rewrite-Regeln für andere Projekte verwenden.

Abstract

Enzymes play a very important role in the metabolism of the organism. All metabolic steps / processes are catalysed by enzymes.
The aim of the master's thesis was to create a database, which is publicly accessible on a website, in order to store graph rewrite rules for chemical, enzymatic reaction graphs and to build on demand chemical reaction networks.
When examining chemical reaction networks, the interest lies in recognizing in which atoms of the educts were converted to which atoms of the products during a biochemical reaction. Graph databases are a useful tool for an efficient and convenient way to store and explore chemical networks. At the current situation networks at the level of atom redistribution cannot be build from databases, because in the databases the reactions are just stored as images and as overall reactions data which is not suited for generative network construction approaches. In addition, reaction databases usually only list compounds and sometimes transformations, not atom mappings itself. The atom mappings of a reaction describe each non-hydrogen atom in a reactant compound and the corresponding atom in a product compound. When the atom mappings of reactions are determined, one can identify reaction centres, which can be used to study reaction mechanisms. The reaction mechanism describes in detail the individual elementary reactions of the entire chemical reaction. It can be clearly seen which reactant atoms have been converted into which product atoms. Thus, the atoms of the chemical reactions within the reaction can be followed. In order to enable this generative approach, the biochemistry must first be made fully accessible as graph rewrite rules, in order to then find the reaction centres, that part of the reaction in which the bonds are broken or newly formed. A chemical graph rewrite rule is an aggregated abstraction of elementary quantum mechanical steps of electron reconfiguration, which must take place in order to implement the structural change of the chemical reaction. In this sense, the context of a graph rewrite rule must include all necessary physical-chemical properties that must be present so that the elementary quantum mechanical steps for implementing the structural change can take place.
These would be the first steps towards the data-driven generation of reaction networks. The aim of this database is to be used by people with little or no chemical experience. A useful application of this approach is the synthetic design of enzyme-catalysed pathways. Because of the network, synthetic pathways are easier to understand and it is also clear which transformation can be catalysed by which enzyme. The automated construction of atom redistribution networks, induced by a pre-specified set of enzymes, required for the interpretation of isotope-labelling experiments in the context of metabolic flux analysis is another future application area of the database. Another advantage would be that the graph rewrite rules are stored in the created database as a minimum and maximum version and can always be accessed. The maximum graph rewrite rule contains all atoms with the atom mappings of the entire reaction. The minimal graph rewrite rule only contains those atoms in which the bonds are directly formed or broken, i.e. the reaction center. In this way the user can use the graph rewrite rules also for other projects from the created database.

Contents

List of Abbreviations

3HP	3-hydroxypropanoate
ADH	Alcohol Dehydrogenase
ALDH	Aldehyde Dehydrogenase
BNF	Boyes Normal Form
BNICE	Biochemical Network Integrated Computational Explorer
BRENDA	BRaunschweig ENzyme DAtabase
CLCA	Canonical Labelling for Clique Approximation
CRN	Chemical Reaction Networks
CSA	Catalytic Site Atlas
DPO	Double Pushout Formalism
DREAM	Determination of REAction Mechanisms
EC	Enzyme Commission
ED	Entner Doudoroff
EMP	Embden-Meyerhof Parnas
ERRD	Enzymatic Reaction Rule Database
GML	Graph Modelling Language
ICMAP	InfoChem-Map
InChI	International Chemical Identifier
InChIKey	International Chemical Identifier Key
IP	Integer Programming
IUBMB	International Union for Biochemistry and Molecular Biology
IUPAC	International Union for Pure and Applied Chemistry
JCBN	Joint Commission for Biochemical Nomenclature
LP	Linear programming
MACiE	Mechanism, Annotation and Classification in Enzymes
MCES	Maximum Common Edge Subgraphs
MCIS	Maximum Common Induced Subgraphs
MCS	Maximum Common Subgraph
M-CSA	Mechanism and Catalytic Site Atlas
MDL	Molecule files
MILP	Mixed-Integer Linear Programming
MØD	MedØlDatschgerl
MWED	Minimum Weighted Edit-Distance metric
RDT	Reaction Decoder Tool
SDF	Structure data files
SMARTS	SMiles ARbitrary Target Specification
SMILES	Simplified Molecular Input Line Entry Specification

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Enzymatic reaction

1.1.1 Enzyme

The role of enzyme in metabolism

Enzymes play an important role as biocatalysts in the metabolism. For instance, cytochrome P450 heme proteins catalyzes reactions such as the biosynthesis of steroid hormones, reactions with cellular macromolecules (DNA, RNA, proteins), they are also involved in the oxidation of unsaturated fatty acids to intracellular messenger substances, in the metabolism of water-insoluble substances by oxidation and lot of more. [1]
Other enzymes are for the digestion of nutrients. Pepsinogens are produced in the stomach, which hydrolyses peptide bonds with phenylalanine, tyrosine and leucine. This helps digest meat with a high collagen content. The enterocytes of the small intestine in turn produce aminopeptidases, carboxypeptidases, endopeptidases and gamma glutamyl transpeptidase, which are involved in the digestion of smaller peptides. [2]
These are just a few examples, enzymes have a multitude of functions in the organism.
The Enzyme Commission (EC) numbers represent enzymatic reactions and also serve as identifiers for enzymes and enzyme genes. The classification of the EC numbers is done manually on the basis of published experimental data on individual enzymes by the Joint Commission for Biochemical Nomenclature (JCBN), of the International Union for Biochemistry and Molecular Biology (IUBMB) and the International Union for Pure and Applied Chemistry (IUPAC). [3] Enzymes are divided into six classes according to their reaction type, which are given by IUBMB. These classes are: oxidoreductases, transferases, hydrolases, isomerase, lyases and ligases. [4]

1.1.2 The mechanism of enzymatic reactions

An enzyme reacts reversible with a substrate at its active center, thus catalyzing the chemical reaction. The enzyme is not consumed and does not change the equilibrium constant of the reaction. But it speeds up the reaction and it lowers the activation energy of the reaction. The substrate is bound to the active center by non-covalent forces, especially Van der Waals forces, hydrophobic interactions, hydrogen bonds or electrostatic forces are the main forces when there is an interaction at the active center of the enzyme with the substrate. The combination of enzyme and substrate is known as an enzyme-substrate intermediate complex. The products are formed, which then dissociate and separate from the enzyme surface and release the enzyme. The enzyme is then ready for another new substrate reaction.

$$\mathrm{E} + \mathrm{S} \underset{\mathrm{k\text{-}1}}{\overset{\mathrm{k_1}}{\rightleftharpoons}} \mathrm{ES} \xrightarrow{\mathrm{k_2}} \mathrm{E} + \mathrm{P}$$

Via the active center, the enzyme reacts reversibly with the substrate to form the enzyme-substrate complex with the rate constant k_1 in the next step with the rate constant k_2, the product and the enzyme, which is unchanged, are formed from the substrate. [5]
Without the enzyme, the product would have to form directly from the sub-

strate, which would be slow due to the high activation energy at the same temperature. The increase in temperature would destroy unstable substrates. There are two ways of interaction between the enzyme and the substrate:

1 The classic key-lock principle

2 And the newer model of induced fit

Key-lock principle:

As the name suggests, the substrate fits into the specific enzyme like a key in its lock. Enzymes are very specific and only catalyze a single or similar substrate. The substrates have a complementary structure/ shape to the enzyme, which fit spatially to one another so that the enzyme can fulfil the biochemical function. [6]
This principle was described by Emil Fischer in 1894. Through non-covalent interactions, a real stable transition state is formed between the enzyme and the substrate, whose real binding strength is called affinity (which can be determined by Km Michaelis Menten Constant).

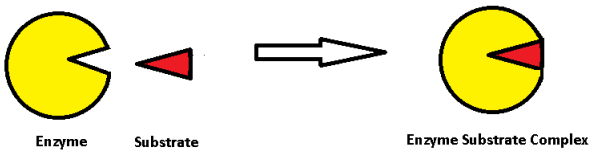


Figure 1.1: Image of the Key-Lock Principle

Induced fit:

The Induced Fit is an extension of the key lock principle and was postulated by Daniel E. Koshland in 1958.
The active center of the enzyme and the binding site of the substrate do not need to be exactly complementary. The active site is flexible and can change its shape/ geometry until the substrate can match and is fully bound. [7]
Thus the enzyme has an initial conformation which attracts the substrate, the enzyme surface is flexible and assuming a change in conformation when the substrate is bound. So, the substrate thus induces a change in shape of the enzyme. [8] As soon as the substrate has docked on the active centre, the conformation of the enzyme changes, which means that the substrate fits better into the active center. Only then does the enzyme-substrate complex form. After the release of the products, the enzyme returns to the original conformation.

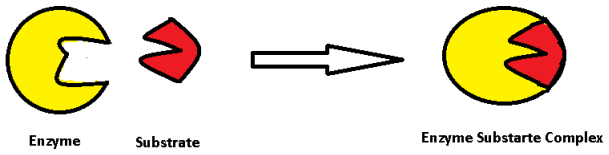


Figure 1.2: Image of the Induced fit Principle

1.1.3 Michaelis menten theory

The Michaelis Menten equation describes the reaction kinetics of an enzymatic reaction. The velocity of an enzymatic reaction depends on the concentration of the enzyme and also on the concentration of the substrate. The higher the concentration of the enzyme, the higher the velocity of the enzymatic reaction. The dependency of the velocity V depends on the substrate concentration (by a constant enzyme concentration). The enzyme activity has a hyperbolic dependency on the substrate concentration. A higher substrate concentration leads to a higher velocity V. First of all, the rate of the reaction velocity increases quite rapidly with increasing the substrate concentration. Until a point is reached where the velocity increases less and less quickly. Finally, a plateau is reached, where the enzymes are completely saturated. So, the maximum velocity $Vmax$ is observed, with which a certain amount of an enzyme E can convert the substrate S into the product P. [9]
The substrate concentration at which $Vmax/2$ is observed is the Michaelis Menten constant Km. Km is an important characteristic of an enzyme, which indicates the relative binding affinity of an enzyme for a specific substrate. The lower Km, the higher the affinity of Enzyme to a substrate. The higher Km, the higher the substrate concentration, so that the reaction takes place at half the velocity at a given enzyme concentration, the lower the affinity of the enzyme for the substrate. If Km is low, a low substrate concentration is sufficient to reach $Vmax/2$ (= so the affinity of the enzyme is very high).

$$\mathrm{V} = \frac{Vmax \cdot [S]}{Km \cdot [S]} \tag{1.1}$$

[1] https://depts.washington.edu/wmatkins/kinetics/michaelis-menten.html

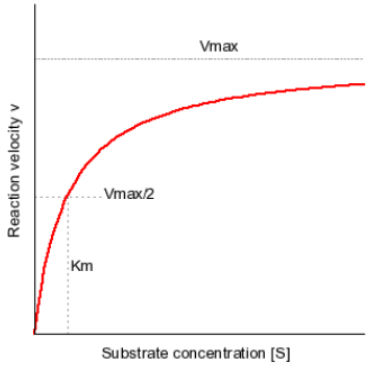


Figure 1.3: The reaction rate of an enzyme reaction depending on the substrate concentration according to the Michaelis Menten equation[1]

1.2 Enzymatic databases

Databases can be suitable tools for biochemical research, since databases store and collect various information about the enzyme, such as the kinetics of the enzyme-catalyzed reaction, the structure or the metabolic function. [4] Other databases store information about certain aspects of enzyme function, enzyme classes, organisms or metabolic pathways. [4]
BRaunschweig ENzyme DAtabase (BRENDA) stores different useful functional information data for all classified enzymes [10] and the metabolism databases MetaCyc have specific metabolic data information, which makes it able to study metabolic pathways for several organisms. [11] [4] Mechanism and Catalytic Site Atlas (M-CSA) in turn shows activity and reaction mechanisms of enzymes. [12] These three important databases are described further in the following section.

1.2.1 MetaCyc

Figure 1.4: MetaCyc Logo[2]

The MetaCyc database is for metabolic data that describes enzymatic reactions, enzymes, small metabolic compounds and metabolic pathways from different organisms that have been experimentally validated and cited from the scientific literature, which is useful for biochemical research, education and so commonly used in different scientific files. (http://metacyc.org/) The metabolic pathways are marked with various comments and details. [11] [13]
MetaCyc also has a large number of atom maps of biochemical reactions from different organisms. These atom maps were determined using the Minimum Weighted Edit-Distance metric (MWED) method. [14] [15] A Mixed-Integer

[2]https://metacyc.org/

Linear Programming (MILP) approach is used to identify the bonds that tend to react. Binding weights are introduced, which represent the tendency for binding changes. During the transformation, costs are determined taking into account the weights and the weight edit distance of a reaction is determined by the sum of the costs of all reaction changes. [16]
The MetaCyc compounds also contain standard Gibbs free energy of formation values. For the same pathway, there do not exist redundant entries. MetaCyc groups together related pathway variations into a common class of pathways. [11]
An advantage of the MetaCyc database is that they use links to other databases (e.g. KEGG, PubChem, ChEBI, ChemSpider, HMDB and lot more). [17] [11]

[3]modified: https://metacyc.org/

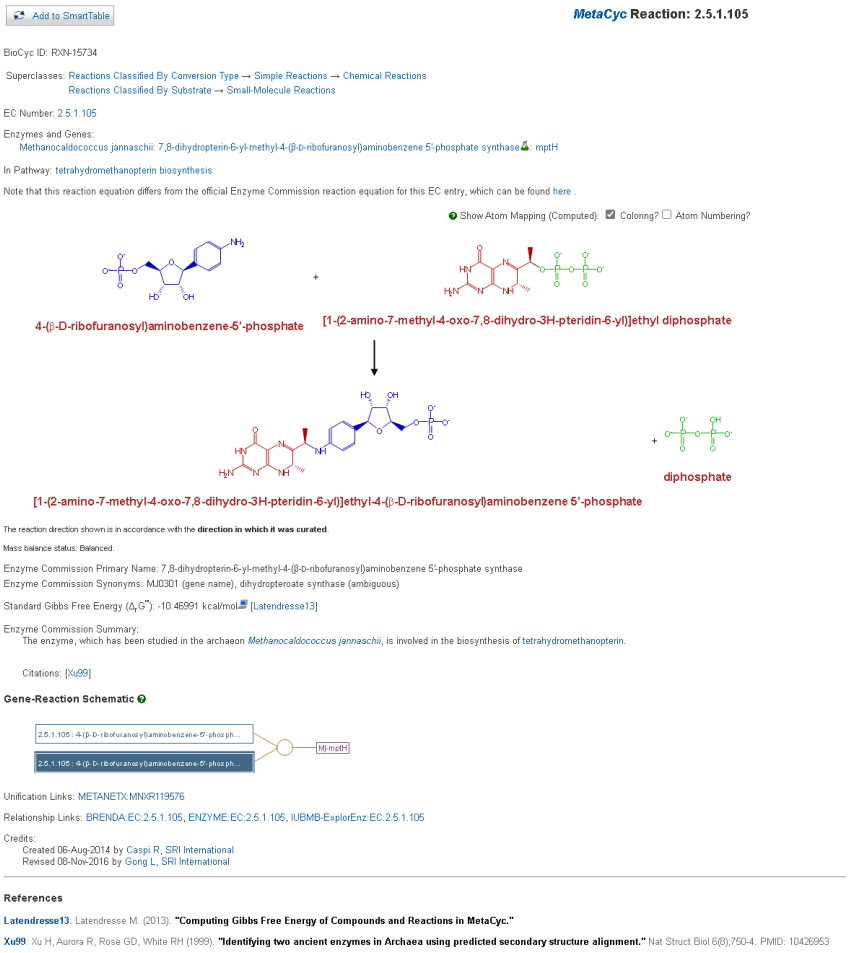


Figure 1.5: The enzymatic reactions are defined in MetaCyc according to the EC number. In this case, the enzymatic reaction 7,8-dihydropterin-6-yl-methyl-4- (-D-ribofuranosyl) aminobenzene 5'-phosphate synthase with the EC number 2.5.1.105 is considered. The page starts with a description of the superclass, the EC number, the name of the enzyme and the pathway. The overall enzymatic reaction is balanced and shown visually with a graphic showing the structure of the compounds. In addition to the atomic map of the educts and products, calculated Gibbs free energies are also written. There are also links, which can be used to lead to external databases. A gene reaction scheme is also visualized, this indicates the relationship between the genes, enzymes and reactions. The literary references are also given on the MetaCyc database reaction page. [3]

1.2.2 BRENDA

Figure 1.6: BRENDA Logo[4]

BRENDA is an enzyme information system, which was created in 1987 at the German National Research Center for Biotechnology in Braunschweig (GBF). (https://www.brenda-enzymes.org/) [10]
The database BRENDA collect and store protein functional data, which contains enzymatic and metabolic data that are used for biochemical research. The metabolic data is extracted from the primary literature and continuously updated. [10]
The enzymes are classified according to the EC system, which was implemented in 1955 by the International Commission of Enzymes. The enzymes are classified based on the chemical reaction they catalyze. [18]
BRENDA has organism-specific information on functional and molecular properties. For instance, information about the nomenclature, enzyme catalyzed reaction, specificity, enzyme structure, enzyme stability, crystallization, organism, ligands, literature references and links to other databases. [18] [10]
The data in BRENDA are stored in a relational database system, where enzymes can be searched by their EC numbers, their names or by the organisms. In addition, the data are structured under "Nomenclature", "Reaction and Specificity", "Isolation and Production", "Function Parameters", "Information on the Organism", "Stability", "Enzyme Structure", "References" and "Application and Technology". [18]
It is also possible to search for ligands, which have an important role in an organism (e.g. substrate/ inhibitor or cofactor/inhibitor). In the BRENDA database the ligands are stored as compound names, Simplified Molecular Input Line Entry Specification (SMILES) string and as Molecule files (MDL). The image of the compound can also be displayed in BRENDA. [10]
BRENDA also allows the calculation of metabolic pathways and the corresponding kinetic data and kinetic values for enzyme–ligand interactions. [10] [18]

[4]https://www.brenda-enzymes.org/
[5]modified: https://www.brenda-enzymes.org/enzyme.php?ecno=1.1.1.1

Figure 1.7: The BRENDA page starts with the EC number and the name of the enzymatic reaction. In this example, the alcohol dehydrogenase with the EC number 1.1.1.1. is displayed. This is followed by the EC tree, which splits into the hierarchical enzyme classification of alcohol dehydrogenase. The structures of the compounds of the overall enzymatic reaction are shown by a graphic illustration. The page provides some information about the enzyme. Such as the synonyms, the reactiontype, the pathway with the associated pathway sources. In addition, external links to other known databases are given in order to obtain even more information. Furthermore, information such as enzyme-ligand interactions, functional parameters, organism-related data, molecular properties, references and so on can be retained.[5]

1.2.3 M-CSA

Figure 1.8: M-CSA Logo[6]

Only the overall reaction is available in most databases. In M-CSA, reactions are resolved in their elementary steps. M-CSA results from the combination of Mechanism, Annotation and Classification in Enzymes (MACiE), which is a database with enzyme mechanisms and Catalytic Site Atlas (CSA), which is a database, that describes the catalytic sites of enzymes. The M-CSA database shows active centres and reaction mechanisms of enzymes, which contains 961 entries with 423 detailed mechanisms and 538 with information on the catalytic sites (www.ebi.ac.uk/thornton-srv/m-csa). [12]
The big advantage is that known enzyme reaction mechanisms and catalytic sites can be searched with the M-CSA database and the data information can be used to understand the enzyme function and its enzyme development. The overall reaction is described in detail as a list of reactants and products containing the EC numbers. M-CSA contains complete catalytic reactions, where each step contains brief descriptions of the chemical mechanism. Information about all cofactors and the catalytic residues is also described. These are supported/ extracted by primary literatures. [12]

[6]https://www.ebi.ac.uk/thornton-srv/m-csa/
[7]modified: https://www.ebi.ac.uk/thornton-srv/m-csa/

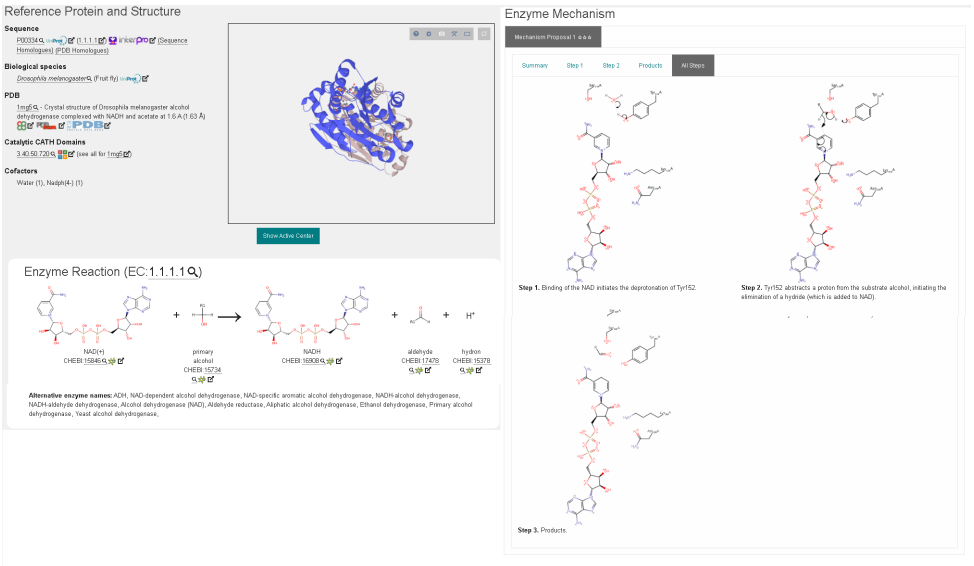


Figure 1.9: The reaction page of the M-CSA database starts with a brief, informative description of the enzymatic reaction. Further links to various databases have been made available in order to obtain more information about the enzymatic reaction. Cofactors and enzyme synonyms have also been given. The enzyme with the active center has been shown in a three- dimensional graphic. The overall enzymatic alcohol dehydrogenase reaction with the EC number 1.1.1.1. is shown below in a further illustration. The educts are NAD(+) and a primary alcohol, which are consumed during the reaction. The products produced after the reaction are NADH, aldhyde and hydron. The overall reaction is also divided into its elementary steps, which have short descriptions of the catalytic residues functions in the reaction and describes the reaction mechanism. These chemical steps are accompanied by electron flow arrows. In the first step, TYR152 is deprotonated. The next step is an elimination of the hydride, which is caused by the uptake of a proton from a primary alcohol by TYR152. The last step shows the products NADH and aldehyde of the alcohol dehydrogenase reaction. References are also given on the M-CSA database. [7]

1.3 Computer representation of molecules

The information of the chemical structure is utilized in similarity searches, reaction retrieval, synthesis planning and drug discovery. [19]
However, the use of trivial names and chemical formula is inadequate, as little or no standardized information about the connectivity of the atoms is available. In order to be able to carry out these applications, a machine readable representation of a molecular structure is necessary to be stored in databases. [20] [21]
For instance, line notation representation and chemical table files, which are simple representation formats of compounds that encode molecular graphs, have been implemented to enable efficient data processing by computers.

1.3.1 Line notation format

Chemical species are represented by line notations, which represent structures as linear strings of characters that can be easily processed by computers. These are particularly popular with chemists and are utilized in cheminformatics. [22]
In the following section a brief overview of the syntax of the most frequently used line notation format SMILES, atom mapping reaction SMILES, SMiles ARbitrary Target Specification (SMARTS), SMIRKS, InChI and International Chemical Identifier Key (InChIKey) are described. These line notation formats are also stored in the Enzymatic Reaction Rule Database (ERRD) that is created in this master thesis and can be retrieved at any time.

SMILES

The SMILES (= Simplified Molecular Input Line Entry System) is a suitable format and most frequently used identifier for storing molecules in databases. [21]
The SMILES specification has a form of a line notation, which describes the structure of chemical species and reactions. [23]
SMILES indicates a molecular structure as a two- dimensional valence- oriented graph. This text- based format was developed to be machine and human readable. The advantages of the SMILES string is that it is compact and the grammar/ syntax of it is easy to understand and to read, because it is similar to chemical structural formula. [22] [24]
The SMILES format can be converted into different useful chemical data by using various chemical toolboxes such as *OpenBabel*. Due to these advantages, it is often used by chemists. A disadvantage of this format is that it is not unique, but if the molecule is canonized, a standardized version of the SMILES can be obtained respectively derived. [24]

SMILES syntax

Every SMILES string consists of a sequence of characters, which ends with a space. Hydrogen atoms may be included implicitly or explicitly. [22] [25] [26]

- Atoms
  By using atomic symbols each non hydrogen atoms can be represented. The atomic symbols are enclosed in square brackets "[" and "]". Not only uncharged atoms can be represented, but also charged atoms. But in some cases, the square brackets may be omitted. For instance, for organic atoms

(B, C, N, O, P, S, F, Cl, Br, and I), for atoms with no chiral centre, for normal isotopes, for atoms that have no formal charge and for atoms that have the number of hydrogens attached implied by the SMILES valence model. [22] [25] [26]

- Bonds
  Bonds are represented by other symbols. "-" stands for a single bond,"=" stands for a double bond and "#" stands for a triple bond. The symbol "-" can be omitted when adjacent atoms are connected to each other by a single or aromatic bond. Ionic "bonds" are not specified directly. Formal charges are written as a disconnected structure by using a dot "." symbol. [22] [25] [26]

SMILES	Chemical Nomenclature	Molecular Formula
CC	Ethane	(CH3CH3)
C=O	Formaldehyde	(CH2O)
C=C	Ethene	(CH2=CH2)
C#N	Hydrogen Cyanide	(HCN)

- Branches
  Branches, which can be nested or stacked, are specified by using parentheses "(" and ")". [22] [25] [26]
- Rings
  For describing a ring in SMILES, one bond must be broken (that leaves a connected acyclic structure). The atoms, where to bond is directly broken, are given a numerical label to indicate the ring.
  Aromatic structures may be specified directly or in Kekulé form with alternating single and double bonds. Or using the aromatic symbol ":" between the atoms of the aromatic ring. Or writing the atoms in lowercase letters, which is mostly used. [22] [25] [26]
- Configuration around double bonds
  With the characters "/" and "\" the configuration around double bonds can be represented. [25] [26]
- Configuration around tetrahedral centres
  In the SMILES syntax chirality specification is based on the order in which the neighbours occur in the SMILES sequence.
  Tetrahedral centres may be indicated by a chiral specification (@ or @@) written as an atomic property following the atomic symbol of the chiral atom. [22] [25] [26]

In order to get a better overview of the SMILES syntax, the molecules (3S,6R)-3-methyl-6-(prop-1-en-2-yl)deca-3,9-dien-1-yl acetate and 4-hydroxy-3-methoxybenz aldehyde are shown in the following images 1.10 and 1.11 with their corresponding SMILES string.

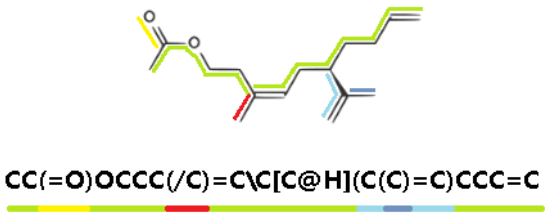


Figure 1.10: Shows both the structural and the SMILES representation of (3S,6R)-3-methyl-6-(prop-1-en-2-yl)deca-3,9-dien-1-yl acetate. The longest path of the spanning tree in the compound are labelled with a green colour and the branches in different colours. The branches are embedded in parentheses. The characters "/" and "\" are used to specify the configuration around the double bond and the configuration around a tetrahedral centre is indicated with the symbol "@".

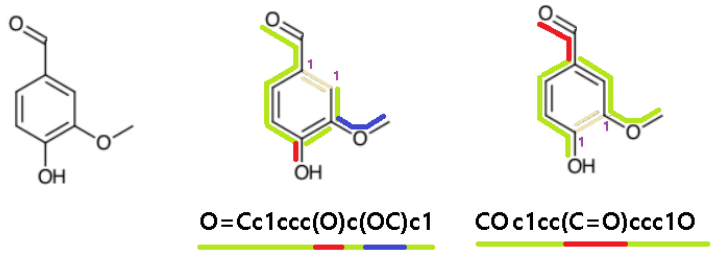
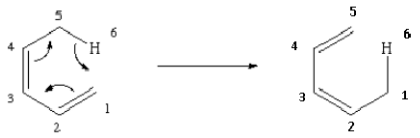


Figure 1.11: Shows both the structural and two possible SMILES representations of 4-hydroxy-3-methoxybenzaldehyde. The same compound does not necessarily have one unique SMILES string as shown in this example with 4-hydroxy-3-methoxybenzaldehyde. Depending on the sequence in which the spanning tree is run, there are different SMILES strings for the same compound. In this example the longest path of the spanning tree has been marked in green and the branches in different colours. The aromatic ring is broken and the number one is indicated on the broken carbon atoms. The aromatic carbons are marked with lower case letters.

Atom mapping reaction SMILES

Chemical reaction can also be described by the SMILES syntax, where the substrate and product molecules are separated by the greater than ("> >") symbol from each other. On the left side the educt molecules and on the right side the product molecules are placed. Multiple molecules on each side can be separated

from each other by using a dot symbol (“.”). [27] [28]
The atom mapping is encoded by using non-negative integer labels between square brackets and the atoms are represent by utilizing atomic symbol letters. The non-negative integers follow after ”:” sign. Two atoms are connected with each other by a single bond if there is no extra symbol between the square brackets. If there is an ”=” or ”#” between two square brackets then these two atoms are connected with each other by a double respectively triple bond. [28]
Each atom is labelled and paired with the corresponding atom of the other side of the chemical reaction. This mean that SMILES atom maps allow specification of reactant atoms with the correspond atoms in products. One reactant and one product have the same atom map. Because of this, it creates a one-to-one mapping between the atoms of the substrates and the products. Without this atom mapping information, it is difficult to recognize the bond changes that occur during the reaction. [27] [28]
In this master thesis the atom mapping reaction SMILES were downloaded from the MetaCyc database and were used as input to parse into the rewrite rules using the automatic generative network construction tool MedØlDatschgerl (MØD). These atom mapping reaction SMILES are stored in the database and also available on the ERRD website.



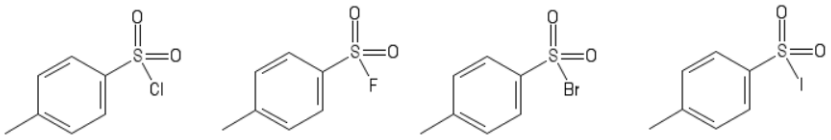
[CH2:1]=[CH:2][CH:3]=[CH:4][CH2:5][H:6]>>[H:6][CH2:1][CH:2]=[CH:3][CH:4]=[CH2:5]

Figure 1.12: A hydrogen rearrangement and the corresponding atom mapping reaction SMILES is illustrated. The atoms are mapped by utilizing integers inside the square brackets. Usually, hydrogen atoms are not labelled and left out, but if a hydrogen atom is directly involved in the reaction, then it is also mapped. In this case the hydrogen with the integer six is connected with the atom, which is labelled with a carbon and has the integer five as mapping. After the rearrangement, the hydrogen is now connected by a single bond with the carbon atom with the integer one. So, with the one-to-one mapping it is possible to trace the atoms through the reaction.[8]

[8] modified: [27]

SMARTS

Another useful line notation is SMARTS (= SMILES arbitrary target specification), which is an extension of the SMILES to determine molecular fragments. [29]
The search for substructures or certain patterns in a molecule is of great importance in chemistry. The SMARTS format describes a molecular pattern and it is developed by Daylight Chemical Information Systems. [30]
For more information look at their manual.
(www.daylight.com/dayhtml/doc/theory/index.pdf) [28]
Finding a specific subgraph, i.e. a pattern in a molecule, can be carried out using the SMARTS language. SMARTS, which is an extension of the SMILES, is a text-based format in which you can specify substructures using rules. SMARTS is often used to find or to filter certain structure groups in a database. Almost all SMILES specifications are valid SMARTS targets. The RDKit library has implemented functions that can compare two molecular SMARTS patterns. [31] [28]
The SMARTS of compounds are also stored in the database and available on the website ERRD.



c1cc(ccc1~[#16](=[#8])(=[#8])-[Cl, Br, F, I])~[#6]

Figure 1.13: All four compounds can be mapped with the same SMARTS pattern. In SMARTS, atomic properties are defined using primitive symbols. This includes various specifications of atoms and charges. [# 6], [# 8] and [# 16] represent a carbon atom, an oxygen atom and a sulfur atom. Due to the specification [F, Cl, Br, I] one of the halogens can sit at the point of the compound. There are also bonding symbols to describe the connection between the atoms. A missing bond symbol means a single bond or an aromatic bond in the molecule. Double bonds are defined with "=" and if any bond could be, it is indicated with the character tilde "˜".

[9] modified: http://efficientbits.blogspot.com/2018/04/rdkit-reaction-smarts.html

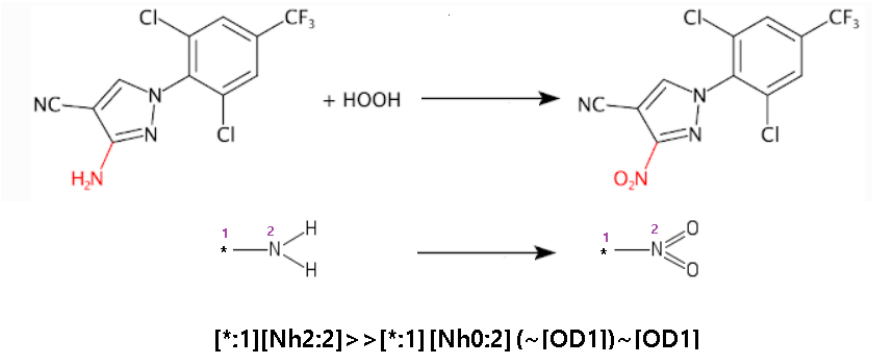


Figure 1.14: A reaction query can consist of a sub-part of the reactant and the product, which are separated by the "> >" symbol. There is a hit if the SMARTS matches within the reaction. This image shows only one example of a reaction that has been mapped by this particular SMARTS pattern, it can match any reaction with an amino reactant and a nitro product. The substructure that applies to this reaction SMARTS pattern is marked in a red colour. The amino group is oxidized to a nitro group during the reaction. The wildcard "*" and the tilde symbol "~" mean that any atom or any bond can follow. [Nh2: 2] and [Nh0: 2] represent an aliphatic nitrogen with two implicit hydrogen atoms respectively with no implicit hydrogen atom. [OD1] stands for a one-connected aliphatic oxygen.[9]

SMIRKS

Smirks is a reaction transform language and is a combination of SMILES and SMARTS to meet the double requirements for a generic reaction, which are the expression of a reaction graph and the expression of indirect effects. It is a limited version of reactive SMARTS in which the atomic bonding patterns change. [28]

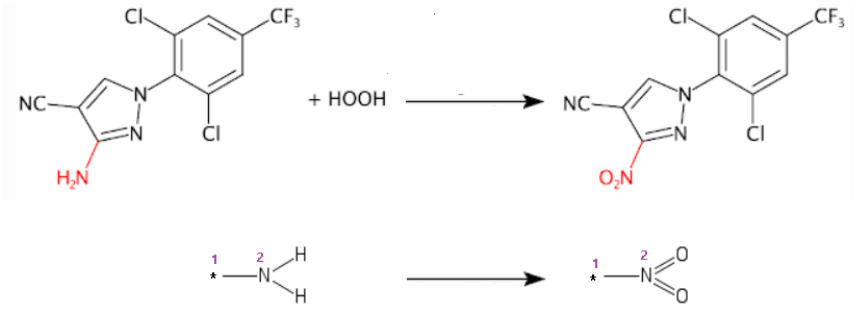


Figure 1.15: The hydrogen atoms that change during the reaction on one side are mapped on the other side of the equation. The bindings representations are SMILES expressions and atomic representations in which the bindings do not change are SMARTS expressions (otherwise SMILES expressions must be used for the atomic expressions). This image shows only one example of a reaction that has been mapped by this particular SMIRKS pattern, it can match any reaction with an amino reactant and a nitro product. The substructure that applies to this reaction SMIRKS pattern is marked in a red colour. The amino group is oxidized to a nitro group during the reaction. The wildcard "*" means that any atom can follow. Since oxidation occurs and the hydrogen atoms are no longer present on the product side, the hydrogen atoms are explicitly given. [ND3: 2] stands for a three-connected aliphatic nitrogen.[10]

Inchi

InChI(= International Chemical Identifier) is a unique line notation, which encode molecular graph information. InChI strings are intended for the use by computers and are also commonly used by chemists. This format was developed to be machine readable. In contrast to the SMILES format the InChI syntax is not suitable to be read by humans. [32]
The algorithm ensures that one molecule is identified by one InChI, which enables a quick comparison of molecules. [33] [21]

[10]modified: http://efficientbits.blogspot.com/2018/04/rdkit-reaction-smarts.html

InChI syntax

The InChI string begins with "Inchi=" and (with the InChI version number) follows by a sequence of layers and sub-layers, with each layer offering one specific type of information. The InChI layers are separated from each other by the symbol slash "/", followed by a lower-case letter (except for the first layer and the chemical formula) [34] [35]

Examples for InChI strings of zealexin B and of morphine are illustrated in the following images 1.16 and 1.17:

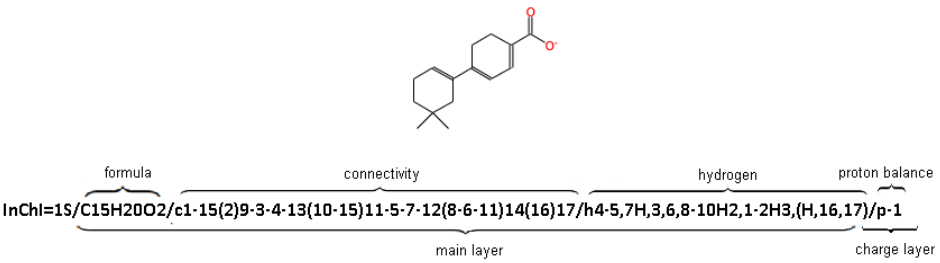


Figure 1.16: The structural formula and the corresponding InChI of the molecule zealexin B is illustrated. In this example, the InChI is divided into the main and charge layers. The main layer consists of the formula, the connectivity, which describes the atomic connectivity between the atoms and the hydrogen connectivity, which defines the connectivity of the hydrogen atoms between the atoms. The charge layer consists only of the proton balance.

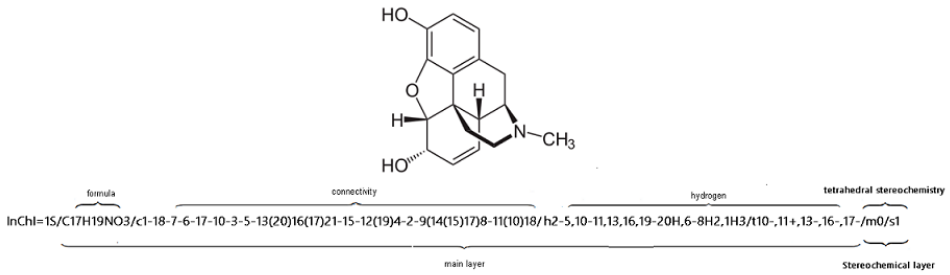


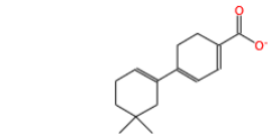
Figure 1.17: The structural formula and the corresponding InChI of the molecule morphine is shown. In this example, the InChI is divided into the main and stereochemical layers. In addition to the main layer, the stereochemical layer is also added, which define the tetrahedral stereochemistry of morphine.

[10] modified: [35]

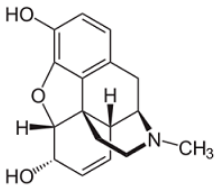
InChIKey

The InChIKey consists of a 27-digit string, which is the compressed version of the InChI. InChIKey leads the search for chemical compounds on the internet and is suitable for database storage. InChIKey consists of 3 parts, where the first 14 lines are based on the connectivity and proton layers. The next 9 characters refer to the other InChI layers. The last part describes the protonation layer. The InChIKey use separators, which are dash symbols. All symbols of InChIKey except the separators are uppercase letters. The InChIKey are also not human-readable. [36]

Examples for InChIKey string of zealexin B and morphine are illustrated in the following image 1.18:



InChIKey=FLTRYYPYDXSVMK-UHFFFAOYSA-M



InChIKey=BQJCRHHNABKAKU-KBQPJGBKSA-N

Figure 1.18: The structural formula and the corresponding InChIKey of the molecule zealexin B and morphine are shown. The 27-digits are clearly recognizable, the first 14 digits represent the connectivity and proton layers, followed by the 9 digits, which characterize the other InChI layers. As well as the InChI, the InChIKey are not human readable.

1.3.2 Chemical table files

A chemical structure file is a text-based chemical file formats that describe molecules and chemical reactions. The line notations SMILES, InChI and InChIKey have no information about the atomic coordinates or the connections between adjacent atoms of a molecule. [21]
Therefore, I would like to mention the chemical table files MDL and Structure data files (SDF) in the next section, which can be used to encode molecular graphs and store graphs in computers.
MDL are text files, which contain structure information for a single chemical species. SDF are text files, which contain a series of MDL files that are joined together. The format represents compounds, which is divided into two main parts. The first part present all atoms and their properties (such as charge and multiplicity) - the second part characterize the bonds within the molecule. [21]
The MDL and SDF connection tables are stored in the ERRD database and can be downloaded by the user from the website that executes the database.

In the following section the structure of the MDL file and the SDF file are described.

MDL Molecule File

A MDL file consists of a header and a connection table block:

1 Header Block

   a) Description
   Title line, Molecule name

   b) Header with timestamp and source program name

   c) Comment line

   b) V2000-compatibility line

2 Connection Table
   A connection table contains information describing the structural relationships and properties of a collection of atoms.

   a) Counts line
   Every connection table starts with the counts line. It specifies the number of atoms, bonds, 3D objects, and Sgroups. And it also indicate the chiral flag setting and the connection table version.

   b) Atom block
   One line per atom: it specifies the x, y, z coordinates of the atom in angstroms, the atomic symbol and any mass differences, charge, stereochemistry and associated hydrogens for each atom.

   c) Bond block
   One line for each bond: it specifies the bond connections between two atoms and any bond stereochemistry and topology.

d) Atom list block
The atom list block identifies the atom (numbers) of the list.

e) Stext block
Structural text descriptor block.

f) Properties block

3 END Block

In the following image the MDL file from L-alanine is illustrated 1.19

[37]

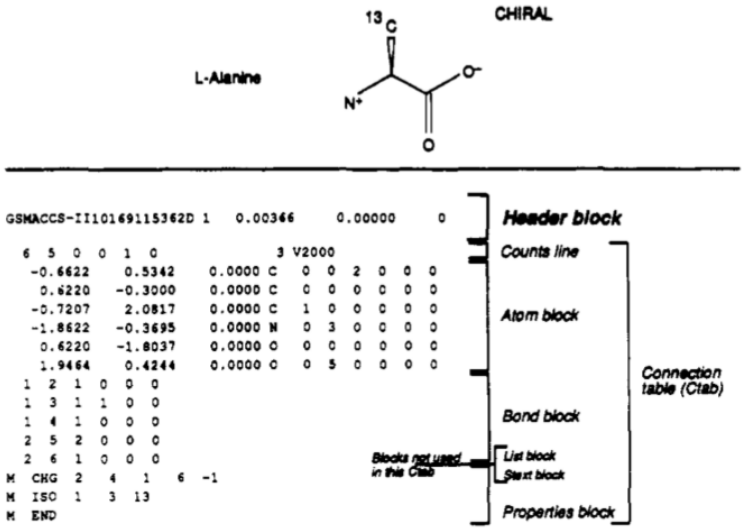


Figure 1.19: This MDL file represents L-alanine. The count line indicates that L-alanine has 6 atoms and 5 bonds and a chiral flag is on, because the fifth number is 1. It has a connection table version V2000. The atom block shows, for each line, one of the molecule atoms, which is represent in x, y and z coordinates. In the fourth column atomic symbols of the atom are listed. The bond block shows the bonds of each atom pair. The first two numbers represent the atoms from the atomic block order. The third column shows the type of bond using numbers. 1 stands for single bond, 2 for double bond, 3 for triple bond and 4 for aromatic bond. The fourth column shows the stereochemistry. In the case of L-alanine, 1 in the fourth column of the binding block means a chirality between the atoms with the number 1 and number 3 which are connected to each other by a single bond. The remaining columns are additional features. The last part is the property block which can contain molecular properties such as charges, radicals, R groups and so on. The molecule ends with the symbol "END".[11]

[11] [37]

SDF Structure Data Files

An SDF (structure-data file) consist of the structural information for one or multiple chemical compounds. SDF contain the MDL. Where each item of a data starts with a greater than symbol ">".
Multiple compounds data blocks are separated from each other by lines consisting of four dollar signs. [37]

1.4 Graph grammar world

1.4.1 Graphs

In chemistry, chemical species and reaction are described with the help of structural formula and reaction mechanism. So, a good way to represent or to model chemical compounds are by graphs and chemical reactions by graph grammar rules.
A graph in general consists of a set of vertices V and edges E, which are connected by two vertices (a source and a target vertices). [38]
Describing molecules as labelled graphs, vertices represent atoms and edges represent bonds. [39]
Graphs can be loaded by using Graph Modelling Language (GML) descriptions or by SMILES format.

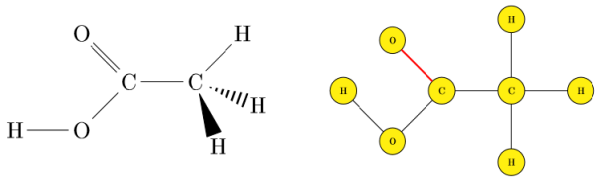


Figure 1.20: On the left side the classic structural formula and on the right side the graph representation of the molecule acetic acid is shown. The yellow nodes represent the atoms and the edges represent the bonds of the molecule. [12]

1.4.2 Graph rewrite rule

With molecules represented as labelled graphs, the mechanism of chemical reaction can be described with graph grammar rules. [40]
This chapter explains the structure of the graph rewrite rule.

Double pushout formalism

A reaction pattern of a rewrite rule or graph transformation rule is described as a Double Pushout Formalism (DPO) [41].
Each chemical rule has the following form:

$$p = \mathrm{L} \quad \xleftarrow{l} \quad \mathrm{K} \quad \xrightarrow{r} \quad R \tag{1.2}$$

A graph production $p : (L \leftarrow K \rightarrow R)$ consists of a production name p and a pair of injective graph morphisms $l : K \rightarrow L$ and $r : K \rightarrow R$. [42]
A DPO transformation rule consists of three graphs L, R and K, which are known as $left$, $right$ and $context$ graphs. The $left$ graph L is a precondition and the $right$ graph R is a postcondition for the application of the graph rewrite rule. The $context$ graph K relates the $left$ and $right$ graph with each other.

[12] https://www.tbi.univie.ac.at/ xtof/Leere/270037/talk01-handout.pdf

Also, it contains two graph morphisms l and r that determine how the *context* is embedded in the $left$ and the *right* graph. [40] [41]
Because there are no particular differences between the $left$ and the *right* graph, the transformation rules are "symmetric". So, with the DPO transformation, the inverse transformation can be defined, which is very useful for the modelling of chemical reactions, because chemical reactions are often invertible. [43]
The inverse transformation has the following form:

$$p^{-1} = \mathrm{R} \quad \xleftarrow{r} \quad \mathrm{K} \quad \xrightarrow{l} \quad L \tag{1.3}$$

So, the DPO formalism is the most "accurate" approach for encoding chemical reactions. [40]
A graph grammar is defined by a set of starting graphs G0 and a set of DPO rules p.

The graph transformation approach consists of looking at a *production* $p : L \rightarrow R$, where the graphs L and R are the left and right sides as a description of an infinite set of direct derivations. [42]
The rule p can be applied to the transformation of a graph G if the $left$ graph L can be found in G, so if a match morphism $m : L \rightarrow G$ can be found. The match morphism m describes how graph L is contained in G. [40]
If a match of $m : L \rightarrow G$ for a *production* exists, it is called a graph homomorphism, in which nodes and edges are mapped from L and G that the graphical structure and the labels are kept. [42]
A new graph H is created by replacing the copy of L with a copy of R, but in such a way that the context K is left intact. G $\xrightarrow{m,p}$ H derivation defines the intermediate graph D and the result graph H as well as the morphisms $d : K \rightarrow D$ and $n : R \rightarrow H$. [40]
The two morphisms d and n describe how the *context* and the *right* graph is contained in D and H graph. The D graph were constructed as a double pushout complement of l and m. The H graph were constructed as a double pushout object of d and r. [40]

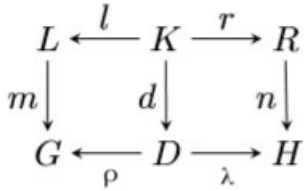


Figure 1.21: Direct Derivation of H from G

An example of a direct chemical derivation using a Diels-Alder transformation rule is illustrated and explained:

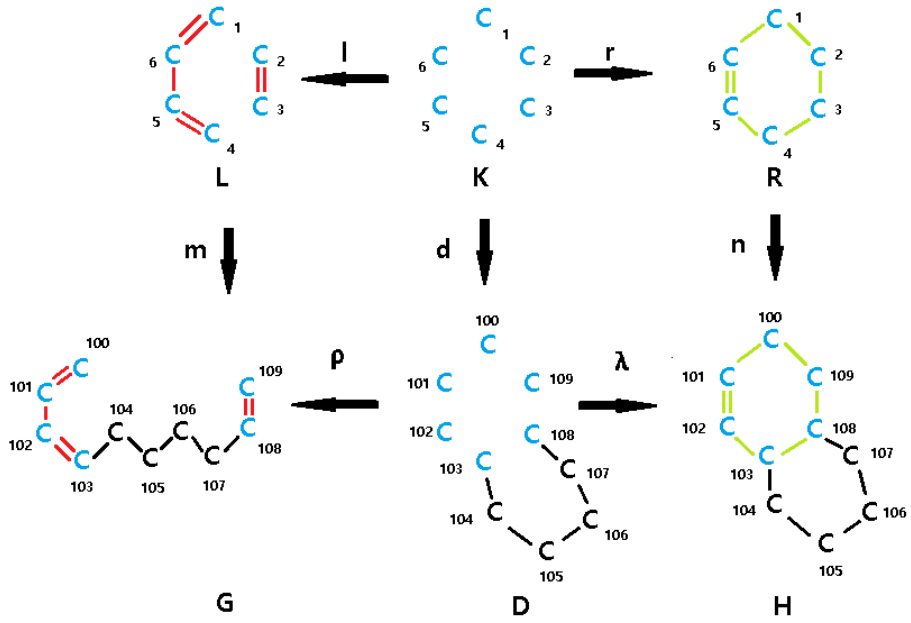


Figure 1.22: The uppercase letters L, K, R, G, D and H represent graph or subgraph objects and the lowercase letters l, m, d, n, ρ and λ represent the graph morphisms. These graph morphisms describe whether a graph is contained in another graph object. The red marked bonds in the left graph L specify the bonds that change during the reaction. The green-coloured bonds in the right graph R are those bonds which have newly formed after the reaction. So, the left graph L is the precondition and the right graph R is the postcondition for the application of the graph rewrite rule. The blue-coloured carbon atoms in the context graph K are those atoms that remain unchanged during the reaction. The host graph G is the actual graph that should be embedded with the transformation rule. The intermediate graph D is a graph in which the precondition L has been removed, but which still contains the context graph K $(D : G \setminus (L \setminus K))$. The result graph H results from the intermediate graph D times the postcondition R less the context graph K $(H : D * (R \setminus K))$.

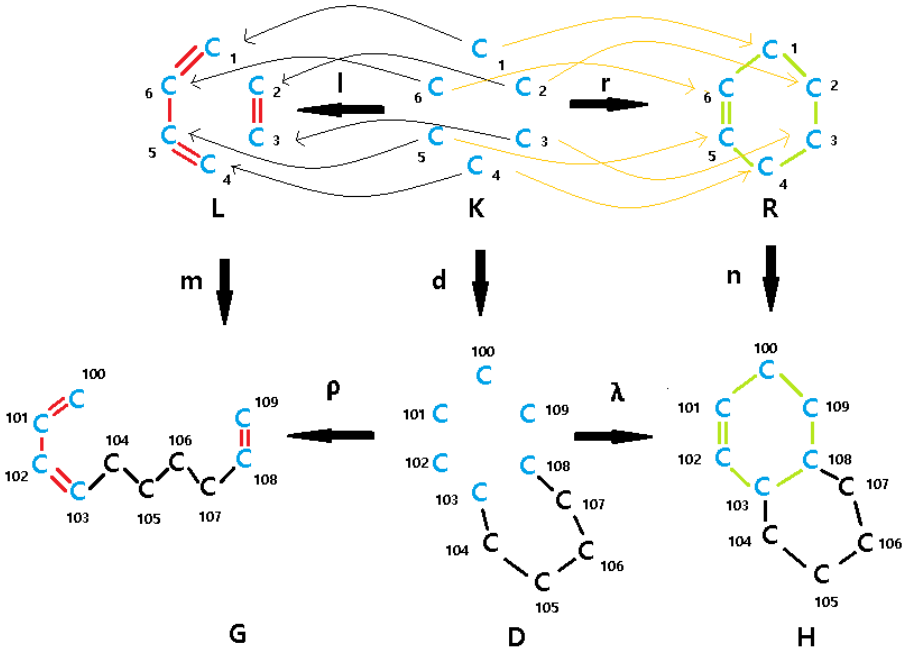
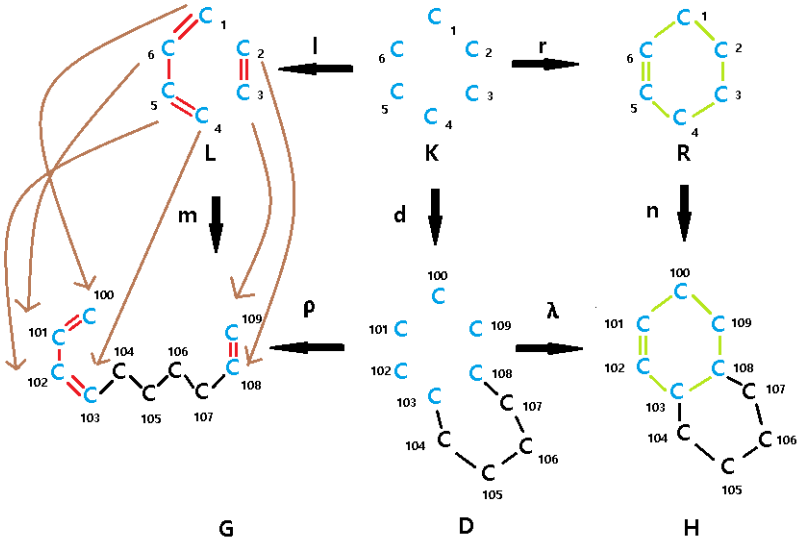
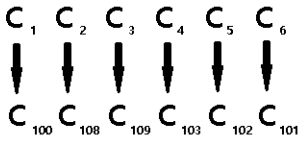
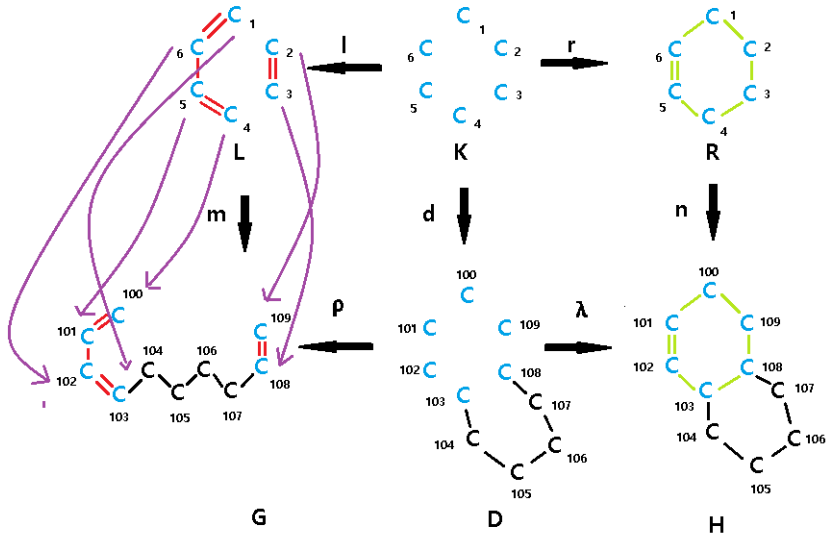


Figure 1.23: The graph morphism l assigns indices to the carbon atoms of the left graph L from the Diels-Alder reaction, which are contained in the context graph K ($L : K \circ l$). This assignment of the carbon atoms is marked with black arrows. The graph morphism r embeds the indices of the context graph K in the product graph R ($R : K \circ r$). This assignment of the carbon atoms is marked with orange arrows. In the case of the Diels-Alder reaction, the context graph K, which do not contain any bonds, are assigned to the left graph L and right graph R, which contain bonds. The left and right graphs are the pre- and post-conditions for the application of the rewrite rule. For instance, in the left graph L, one receives the information that a double bond must be found between the two atoms C1 and C6. This double bond is later removed and replaced by a single bond between C1 and C6 in the right graph R.


l
r
L
K
R
m
d
n
ρ
λ
G
D
H


C1 C2 C3 C4 C5 C6
C100 C108 C109 C103 C102 C101


l
r
L
K
R
m
d
n
ρ
λ
G
D
H

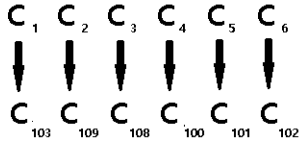

C1 C2 C3 C4 C5 C6
C103 C109 C108 C100 C101 C102

Figure 1.24: The graph morphism m is a subgraph isomorphism, where the left graph L should be found efficiently in the host graph G ($G : K \circ l \circ m$). Thus, the transformation rule can be applied to a host graph. The left subgraph L embeds in the molecule G. Several permutations are possible, which are shown with the brown as well as the purple arrows. With the graph morphism m, a translation is obtained which maps the indices from the rule to the indices in the host graph G. This mapping is obtained in an efficient manner. For a better understanding, the different carbon atom mappings between the left graph L and the host graph G, which are obtained from the graph morphism m, are entered under the two diagrams.

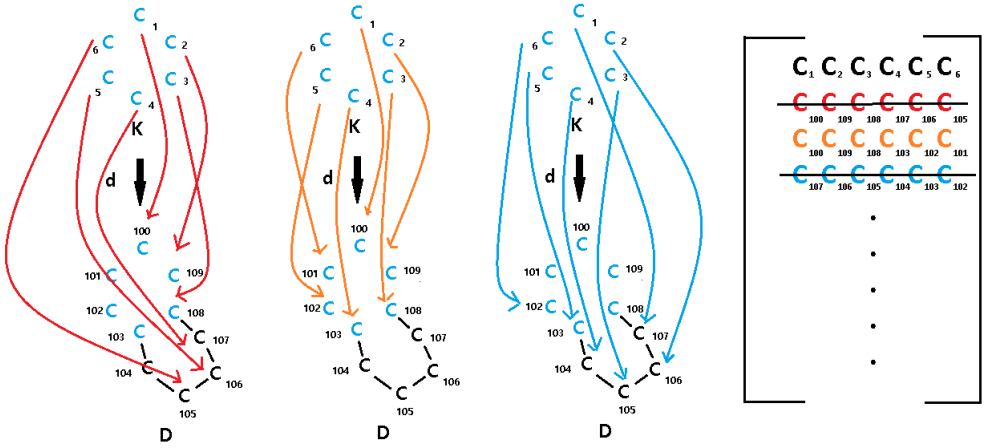


Figure 1.25: Furthermore, an inefficient route can also be taken, which leads to the same result. In the intermediate graph D the precondition has been removed, but the context is still included ($D : K \circ d$). To get the graph morphism d, the carbon atoms of the context graph K are mapped with the carbon atoms of intermediate graph D. This d mapping leads to many embedding possibilities of carbon atoms, since with m atoms of a query and n atoms of a host molecule the mapping possibilities with $n!/(n-m)!$ increase exponentially. So, a list of different embeddings, most of which are incorrect, is given. The inverse mapping ρ can now be used to check the mapped atoms for the precondition. From the exponential list one removes all possibilities which cannot make the correct condition pattern. This leaves only the mapping that has been efficiently calculated using the subgraph isomorphism m. The same result can be found using two different mapping paths, so there is a commutative diagram. Commutation means that regardless of which path is chosen for successive assignments, the list of remaining embeddings must be the same. Both efficient and inefficient routes are thus possible.

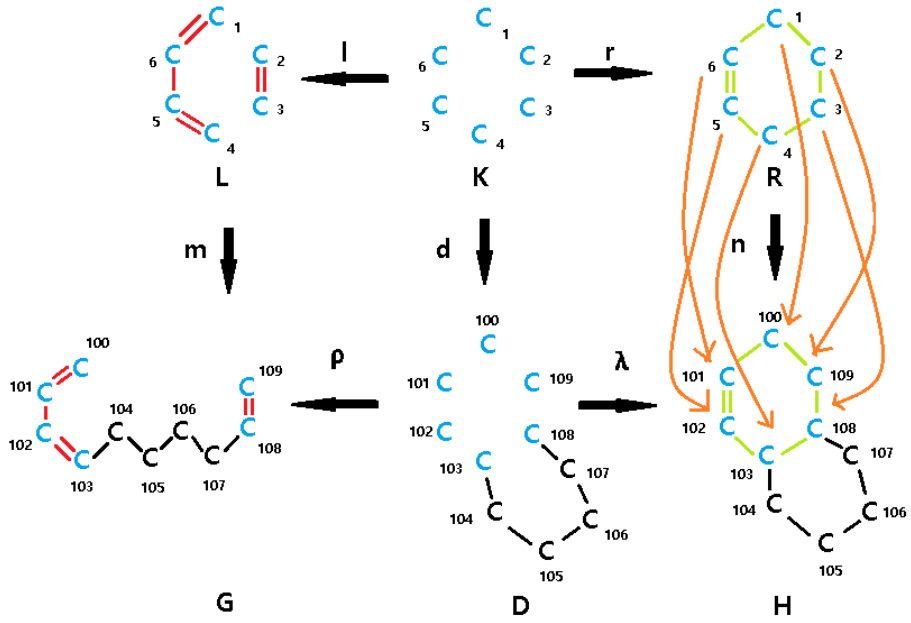


Figure 1.26: The precondition (the binding pattern) has been removed from the host graph G, and so the intermediate graph D is created. Since you have the atom assignments, you can remove exactly the specific bonds. If the postcondition is applied to the intermediate graph D, one arrives at the result graph H. So, if the graph rewrite rule is applied, then the result graph H is generated.

Graph modelling language

The graph modelling language GML format is used to encode the graph grammar rules. The GML is a simple and flexible representation for graphs, which uses the ASCII standard. The structure of the GML format consists of hierarchical key-value pairs. The keys are usually strings and values are usually integers, float numbers, strings, or another key-value list. Here the lists must always be enclosed in opening and closing square brackets ("[" and "]") in GML. [44]

The general structure of the GML representation is illustrated in the following picture 1.27:

```
key1 [
 key2 value2
 key3 [
  key4 value4
  key5 value5
 ]
 key6 value6
]
```

Figure 1.27: *GML* Structure[13]

The first and the third key have a list as value. The remaining keys (key 2, key 4-6) are key-value pairs, where the values can be integers, float numbers or strings.

The Boyes Normal Form (BNF) notation in GML is in shown in the image 1.28

```
gml       ::= keyvalues
keyvalues ::= keyvalue (keyvalue)*
keyvalue  ::= key value
key       ::= ['a'-'z''A'-'Z']['a'-'z''A'-'Z''0'-'9']
value     ::= real | integer | string | list | operator
real      ::= sign? digit  '.' digit+ mantissa?
integer   ::= sign? digit+
operator  ::= '<' | '=' | '>' | '!'
string    ::= '"' instring '"'
list      ::= '[' keyvalues ']'
sign      ::= '+' | '-'
digit     ::= ['0'-'9']
mantissa  ::= ('E' | 'e') sign? digit+
instring  ::= ASCII-{'&', '"'} | '&' ['a'-'z''A'-'Z']  ';'
```

Figure 1.28: BNF notation in GML[14]

[13] [44]
[14] [44]

For modelling a graph rewrite rule a key rule must be specified, within the list of the rule key, four keys must be specified, a string value key is used to define/ name the graph rewrite rule and three list of value keys that specify the three subgraphs ($left$, *context* and *right* subgraph) of a graph rewrite rule. [44]

In the graph modelling language, graphs and graph rewrite rules include the node and edges keys. Within the list of the node, unique IDs are defined, which indicate each atom of the reaction. The atoms are also labelled to define their atom types (for instance, "C"," N" ,"O" and a lot of more atom labels). Within the list of the edge, the nodes IDs are used to specify the source and the target of a bond. The bonds are also labelled to define their bond types (for instance, "-" for a single bond, "=" for a double bond and "#" for a triple bond). [44]

The $left$ subgraph of the rule expresses the local state of a molecule before applying the reaction rule. Within the list value of the $left$ subgraph all edges are added, which vanish during the reaction. Also, all nodes are added in the list, which changed their label during the transformation. So, all edges and nodes are specified within the list value of the $left$ key, which are present in the educt molecules, but absent in the product molecules. The educt molecule graphs are defined. [44]

The *right* subgraph of the rule expresses the local state of a molecule after applying the reaction rule. Within the list value of the *right* subgraph all edges are added, which newly form during the reaction. Just like in the $left$ subgraph, all nodes are added, which changed their label during the transformation. So, all edges and nodes are specified within the *right* value of the *right* key, which are absent in the educt molecules, but present in the product molecules. The product molecule graphs are defined. [44]

The *context* subgraph relates the *right* and the $left$ subgraph to each other and encodes the invariant part of the reaction centre. Within the list value of the *context* subgraph, all nodes and edges are added, which do not change during the reaction. [44]

An example for a DPO and the corresponding graph rewrite rule of the enzymatic reaction catechol 2,3-dioxygenase are shown in the images 1.29 and 1.30.

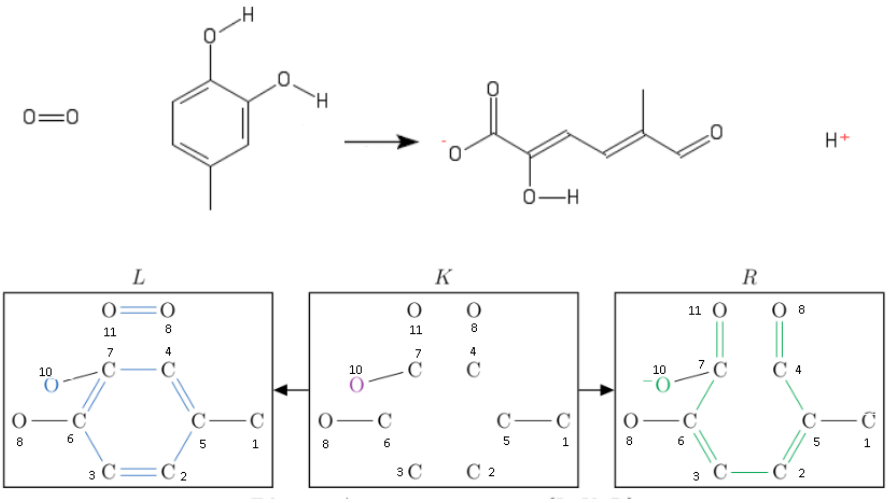


Figure 1.29: In the left subgraph L all bonds and atoms of the compounds oxygen and 4-methylcatechol, which change during the reaction are labelled in a blue color. Within the context subgraph K, all bonds and atoms are added, which do not change during the reaction. In the right subgraph R all bonds and atoms of the compound (2Z,4E)-2-hydroxy-5-methyl-6-oxohexa-2,4-dienoate, which are newly formed, are coloured in green.

```
rule [
        ruleID "Catechol 2,3 Dioxygenase"
        left [
                node [ id 10 label "O" ]
                edge [ source 2 target 3 label "=" ]
                edge [ source 3 target 6 label "-" ]
                edge [ source 6 target 7 label "=" ]
                edge [ source 8 target 11 label "=" ]
                edge [ source 2 target 5 label "-" ]
                edge [ source 4 target 5 label "=" ]
                edge [ source 4 target 7 label "-" ]
        ]
        context [
                node [ id 1 label "C" ]
                node [ id 5 label "C" ]
                node [ id 2 label "C" ]
                node [ id 3 label "C" ]
                node [ id 6 label "C" ]
                node [ id 9 label "O" ]
                node [ id 7 label "C" ]
                node [ id 4 label "C" ]
                edge [ source 1 target 5 label "-" ]
                edge [ source 6 target 9 label "-" ]
                edge [ source 7 target 10 label "-" ]
                node [ id 8 label "O" ]
                node [ id 11 label "O" ]
        ]
        right [
                node [ id 10 label "O-" ]
                edge [ source 2 target 3 label "-" ]
                edge [ source 3 target 6 label "=" ]
                edge [ source 6 target 7 label "-" ]
                edge [ source 7 target 11 label "=" ]
                edge [ source 4 target 8 label "=" ]
                edge [ source 2 target 5 label "=" ]
                edge [ source 4 target 5 label "-" ]
        ]
]
```

Figure 1.30: The rewrite rule of the enzymatic reaction catechol 2,3-dioxygenase is illustrated. Within the key rule list, the name catechol 2,3 dioxygenase is initially defined as a string value key. This is followed by three more value keys that define the subgraphs. Within the value list of the $left$ graph seven bonds and one oxygen atoms are listed, which are converted after the reaction. Within the value list of the *contex* graph, three bonds and ten atoms are entered, which remain unchanged. In the value list of the *right* graph, seven new bonds and one negatively charged oxygen atom are inscribed.

1.4.3 Chemical reaction networks

For a better understanding of chemical processes/ pathways the analysis of chemical reaction networks is essential. [45]

A chemical reaction is defined as follows:

$$aA + bB + \cdots \rightarrow xX + yY + \ldots \tag{1.4}$$

A chemical reaction network Chemical Reaction Networks (CRN) is defined as a set of chemical species that are connected to each other by reactions. Each chemical reaction transforms a set of substrate molecules into a set of product molecules. [45]
A suitable way to represent CRN are directed hypergraphs. [45] [46]

Directed hypergraphs

A hypergraph is a graph, which is made up of a set of V of vertices and a set E of hyperedges $H = (V, E)$, where each hyperedge $e \in E$ may contain any number of vertices and is thus defined as a subset of V. So, a hypergraph is a graph, where an edge can connect more than two vertices. Each hyperedges $e \in E$, also called hyperarc, is an ordered pair of (t(e),h(e)). [45]
This allows to connect several start nodes (tail t) with several end nodes (head h). The tail of a hyperarc refers to the reactants, while its head identifies the products. In chemical reactions networks, the vertices/ nodes describe the chemical species and the hyperedges represent the chemical reactions, which connect the educts and products. With directed hypergraphs, the directions of each hyperedge is assigned, so the connection of the start nodes and the end nodes are defined. [47]

In the following figure 1.31 a hypergraph is illustrated:

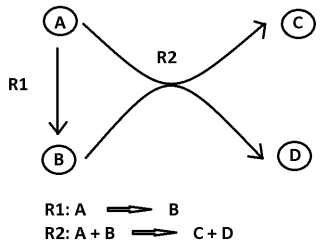


Figure 1.31: Hypergraph: This directed hypergraph consist of two reaction (R1 and R2) with four chemical species (A, B, C, D). Each reaction is represented by a single directed hyperedge connecting educts with products. Hyperarc R1 has t(R1) = (A) and h(R1) = (B), hyperarc R2 has t(R2) = (A, B) and h(R2) = (C, D).

Directed hypergraphs can be illustrated as directed substrate graphs or as bipartite graphs.

In the following figure 1.32 a directed substrate graph is illustrated:

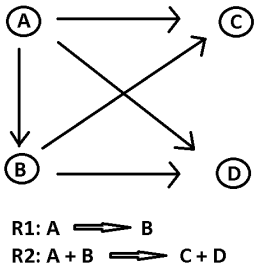


Figure 1.32: Directed Substrate Graph

Bipartite directed graphs, which chemical reaction networks can be described, represent both reaction and metabolites as two different types of nodes. Direct edges just exist if a reaction transforms a metabolite, then the educt nodes are connected with the reaction node, or vice versa. Also, the bipartite graph reflects the original information from the hypergraph. [47]

In the following figure 1.33 a directed bipartite graph is illustrated:

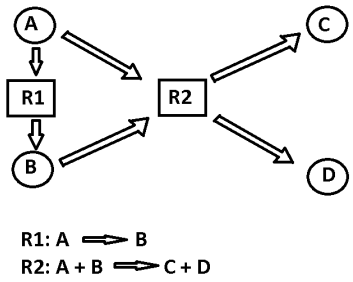


Figure 1.33: Directed bipartite graph: Reaction nodes are drawn as squares and species nodes as circles. This directed bipartite graph consist of two reaction (R1 and R2) and four chemical species (A, B, C, D). The reaction R1 transforms educt A to product B and the reaction R2 transforms the educt molecules A and B to the product molecules C and D.

1.4.4 Causal analysis

Another important advantage of rule-based models lies in their suitability for causal analysis that takes the logically nature of interactions. [48]
Two rules have a causal relation if the first one generates or destroys a subgraph, which is necessary for the second rule to be able to match. The subgraph of the first reaction does not have to be the whole molecule, but only the part that is necessary for the second reaction.
With the notion of causality, the order of reaction in a pathway can be investigated. So, the question can be asked which rules have to happen strictly before other ones or concurrently to other ones, such that the pathway performs the overall transformation. So, it can be asked for any network or pathways, are the reactions in the right order or can they be reordered in such a way that the overall transformation is still the same.

Here I would like to show an example to illustrate the principle of the causal analysis:

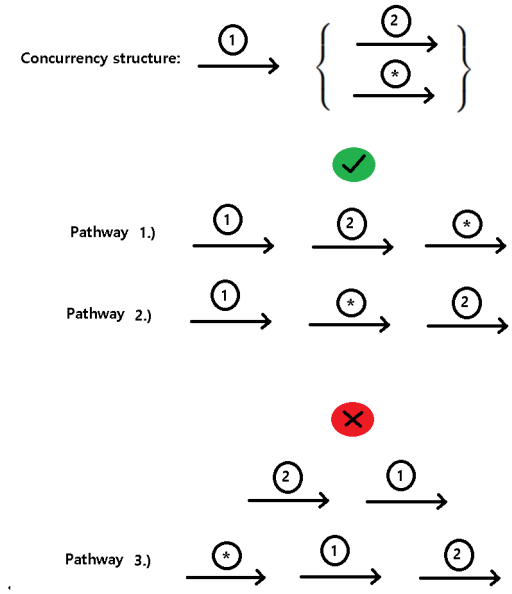
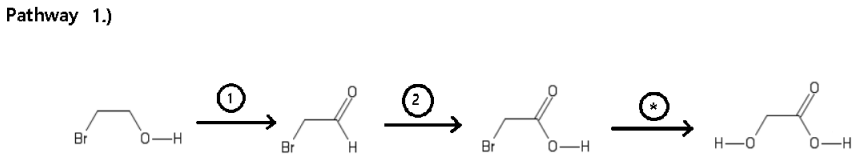


Figure 1.34: On the basis of the subgraph relation, one can define that the enzymatic reaction 1 generates the precondition for the enzymatic reaction 2, but that any parallel transformation reaction * that is not dependent on reaction 1 and 2 can take place between the two reactions. This means that reaction 2 is causally linked to reaction 1. Reaction 1 has to take place before reaction 2. A pathway cannot be constructed where reaction 2 starts first and then reaction 1 follows. The concurrency theory defines which reactions can follow sequentially and which can follow in parallel. The transformation reaction * can react in parallel at any time, such as before reaction 1, after reaction 2 or between reaction 1 and reaction 2. In summary, the concurrency structure says that reaction 1 must take place before reaction 2. The transformation reaction * can run parallel between the two reaction 1 and 2.

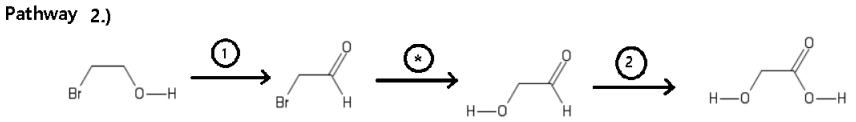


Reaction 1: $BrCH_2CH_2OH + NAD^+ \overset{ADH}{\longleftrightarrow} BrCH_2CHO + NADH + H^+$

Reaction 2: $BrCH_2CHO + NAD^+ \xrightarrow{ALDH} BrCH_2COOH + H^+$

Reaction *: $BrCH_2COOH + OH^- \xrightarrow{Enzyme} OHCH_2COOH + Br^-$

Figure 1.35: Pathway 1: When alcohol is broken down, Alcohol Dehydrogenase (ADH) produces acetaldehyde. In the next step, the acetaldehyde is converted into the acetic acid by the enzyme Aldehyde Dehydrogenase (ALDH). In this example the aldehyde group is formed, which is not there before, this is a fresh subgraph, which is needed for the second rule to match and to be able to perform. Reaction 2 must strictly follow after reaction 1, the reactions must be in this order, it cannot be removed or changed. These reactions 1 and 2 are causally related to each other. In the next step, bromine is exchanged for hydroxide ion and the product of the first pathway is the smallest alpha-hydroxy acid glycolic acid.

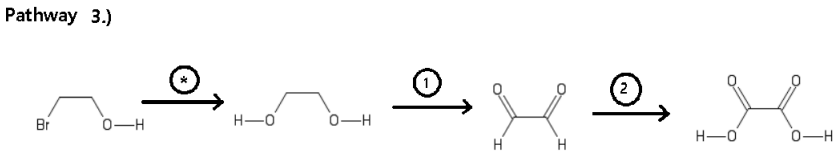


Reaction 1: $BrCH_2CH_2OH + NAD^+ \overset{ADH}{\longleftrightarrow} BrCH_2CHO + NADH + H^+$

Reaction *: $BrCH_2CHO + OH^- \xrightarrow{Enzyme} OHCH_2CHO + Br^-$

Reaction 2: $OHCH_2CHO + NAD^+ \xrightarrow{ALDH} OHCH_2COOH + H^+$

Figure 1.36: Pathway 2: The alcohol is converted into aldehyde due to the reaction 1 with the enzyme ADH. By substituting bromine for hydroxide ions, the intermediate product glycolaldehyde is created through the transformation reaction *. The glycolaldehyde now forms the subgraph for reaction 2 so that it can take place. The aldehyde is converted into the same alpha-hydroxy acid product as in pathway 1. Reaction 1 has to start strictly at the beginning, then the transformation reaction * can follow in parallel, since it gives the same product for both path variants 1 and 2. In the pathways 1 and 2 there is no dependency of the transformation reaction * on the other reactions.



Reaction *: $BrCH_2CH_2OH + OH^- \xrightarrow{\text{Enzyme}} OHCH_2CH_2OH + Br^-$

Reaction 1: $OHCH_2CH_2OH + NAD^+ \xleftrightarrow{\text{ADH}} CHOCHO + NADH + H^+$

Reaction 2: $CHOCHO + NAD^+ \xrightarrow{\text{ALDH}} COOHCOOH + H^+$

Figure 1.37: Pathway 3: But if the transformation reaction * starts first in the pathway, the end product is not glycolic acid as in the first two pathway variations. After the substitution step, ethanediol is obtained. In reaction 1, ethanediol is converted into glyoxal by the ADH enzyme. Since reaction 2 with the enzyme ALDH is causally linked to reaction 1 and the aldehyde is the required subgraph, the end product of pathway 3 will be oxalic acid. The concurrency structure shows that when a certain molecule as the starting compound is given and a certain end product should arrive, then there are clear paths left that describes this reaction path.

1.4.5 Atom maps

A useful application of this approach is in the synthesis planning of enzyme catalyzed pathways. Because of the network, synthetic pathways are understandable, and it is also clear which transformation can be catalyzed by which enzyme.

In a chemical reaction, a set of chemical substances is converted from one compound to another. Since in the course of a reaction the broken and formed bonds respectively the change in the order of the bonds are of great importance, those atoms and bonds play an important role that are directly involved in this rearrangement of electrons and these form the so-called reaction center, which is in a great interest for a chemist. Reaction database usually list only compounds and sometimes transformation, but not atom maps itself. The atom mapping of a reaction describes for each non-hydrogen atom in a reactant compound and its corresponding atom in a product compound. If the atomic correspondence is determined, one can identify reaction centres that can be used to investigate reaction mechanisms. The reaction mechanism describes in detail the individual elementary reactions of the overall chemical reaction. So, it can be clearly seen which educt atoms have been converted to which product atoms and this allows to track the atoms of the chemical reactions.
Experimentally, the atom mappings are obtained through the isotope labelling experiments. The atoms in which a bond change is expected are replaced by isotopes. The positions of these isotopes can be determined using certain techniques (NMR, MS). Thus, atom-to-atom mapping relationships between educts and products are obtained. With the help of the atom-to-atom mapping, the changed part is determined as the reaction center. The isotope labelling experiments give the atomic correspondence and ideas for the reaction mechanism. The major disadvantage with these isotope labelling experiments is that this approach is very expensive and time consuming. Besides, this can hardly be done for all atoms.

After isotope-labelled substances are added to an organism and redistributed it is possible to find out which reaction were possible and which were not possible. And for that you have to be able to track the atoms in the reaction networks. This is currently only possible on small, hand- built networks.

Isotope labelled experiments in glycolysis are commonly used to analyze the activity of the different pathway variations.
It is known that the Embden-Meyerhof Parnas (EMP) and Entner Doudoroff (ED) pathways lead to different carbon traces, but since atom maps are usually not available in databases it can be very exhausting to analyze trace data manually.
With chemistry model based on DPO transformation rules enable the automatic inference of atom traces for complete pathways.
Labelling experiments in glycolysis are commonly used to analyze the activity of the different pathways [49] and in the following image the EMP and ED pathways are illustrated (just carbon atom trace) 1.38:

[15] modified [49]

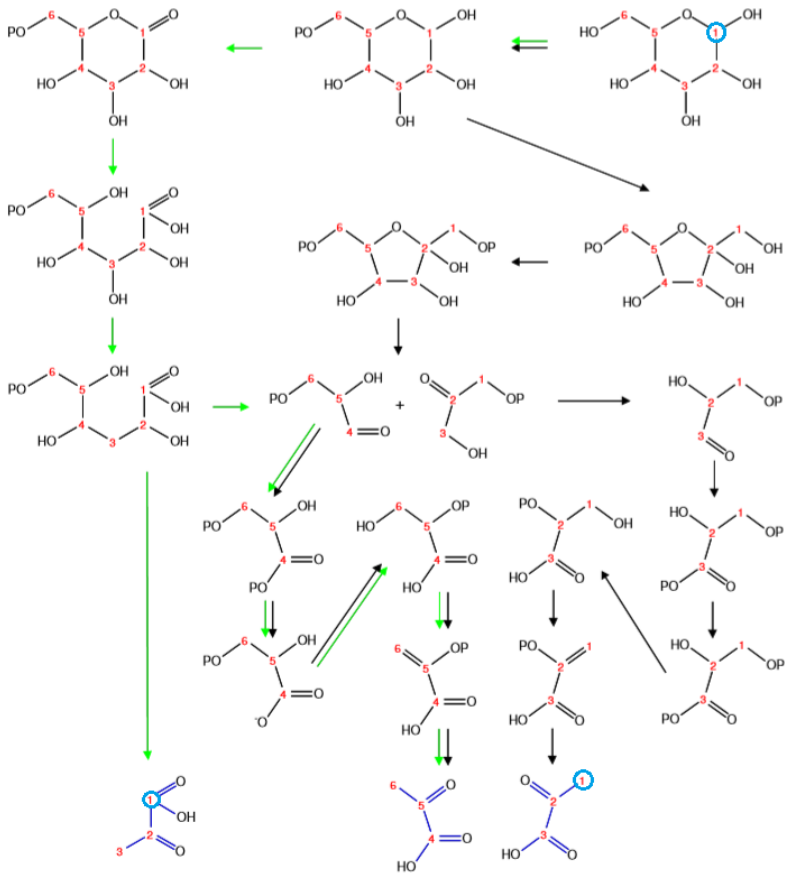


Figure 1.38: Embden-Meyerhof Parnas EMP and Entner-Doudoroff ED pathway[15]

The black reaction arrows illustrate the EMP pathway and the green arrows demonstrate the ED pathway. The carbon atoms are labelled/ mapped with numbers, so to be able to track the atoms during the two reaction pathways. The six carbon atoms from glucose are transformed into two different pyruvate molecules, depending on which of the two pathways were used to transform glucose. The carbon atoms of the created pyruvates are arranged differently. [49]
During the EMP pathway transformation a pentose ring from a fructose 1,6 bisphosphate intermediate is cleaved. In the ED pathway transformation the hexose ring of glycose 6-phosphate is cleaved. The EMP path forms two pyruvate molecules. Marked in the picture with a blue circle, the carbon atom with the label 1 of the pyruvate molecule in the EMP pathway is transformed in one of the two pyruvate products. The carbon atom is positioned very differently from the carbon marked with the label 1 in the ED pathway, which is also marked with a blue circle. [49]
In this way, important information about the mechanism of a reaction can be drawn from the atom mapping. This is a key advantage in application to

large networks, because atom maps make it possible to follow individual atoms through complex CRN. [50] [51]
The representation of chemical reaction based on transformation rules is therefore sufficiently detailed to support the analysis of isotope labelled experiments.

Issue of atom mapping

One big issue is that one would like to have the atom maps, but cannot guarantee that one will get the chemically correct ones. In the MetaCyc database, for example, the atom maps are only calculated and do not guarantee that they are chemically correct.

The following image 1.39 shows a good example of two possible atom maps of the Diels-alder reaction. In this reaction, an unsubstituted diene reacts with an alkene to form a ring of six carbon atoms. Now, it can be questioned, whether the double bond of cyclohexene comes from ethene or one of the two double bonds from diene.

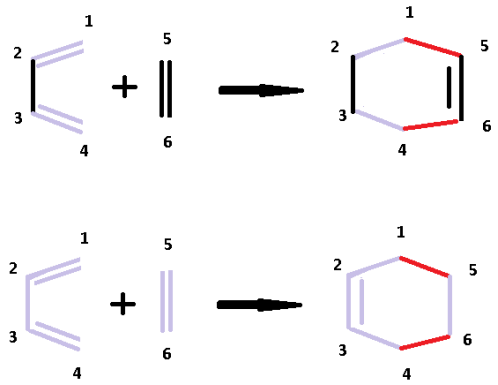


Figure 1.39: Two different atom maps of the Diels-alder reaction are shown. The second atom map is the chemical correct one. Bonds, which are labelled in red are newly formed and bonds, which are labelled in blue are changed in bond order.

In the correct mechanism it leads to a rearrangement of three pi-electron pairs, which form two new sigma-bonds between a dienophile and a diene as well as a double bond. If the atom mapping is correct then 6 bonds change in the Diels-alder reaction and if you have the chemically incorrect atom mapping then you only have 4 bond changes. This is how mistakes come in when you want to trace atoms with chemical incorrect atom maps through chemical reaction networks.

An alternative approach for the identification of atom-to-atom mappings and reaction centers and thus to the reaction mechanisms would be with different atom mapping algorithms. [52]

The following paragraph explains the different approaches and algorithms that compute atom-to-atom mapping for reactions.

1.5 Computer based approaches for atom mapping

1.5.1 Fragment-assembly-based methods

Lynch fragment-assembly-based methods: In fragment assembly-based methods, the reactions are fragmented into small fragments. After all fragments are processed, the fragments on the left side are compared to the fragments on the right side of the reaction to determine which are identical. Identical fragments are removed in the next step, because these fragments are not a part of the active reaction center. The remaining fragments are merged in the educt and product in order to identify the active reaction center. However, the disadvantage here is that it is impossible to obtain the exact position of the reaction site within the parent structure, since ambiguities occur in the fragmentation step. [53] [54]

1.5.2 Common substructure-based methods

With the common substructure-based method, molecules are shown as graphs and all common substructures between the two graphs, i.e. the educt and the product, should be found. The degree of similarity between the graphs can be determined using graph isomorphism respectively the maximum common graph isomorphism. If molecules are the same, i.e. if there is a one-to-one relationship between all atoms and bonds, then this is called isomorphic. So two graphs are isomorphic if there is a one-to-one correspondence between the vertices and the associated edges between the two graphs $G1$ and $G2$. If the molecules are not exactly, but the molecules share a common substructure, then a subgraph isomorphism between the two molecular graphs is given. [55] Molecules represented as graphs, which can be compared with graph isomorphism, play an important role in the chemistry and biology, especially when docking protein ligands, searching in database, modelling the reaction, interpreting molecular spectra and pattern recognition. [33] [55] Two graphs can have more than one common substructure. But the largest common substructure between two graphs that represent molecules are called the maximum common substructure (Maximum Common Subgraph (MCS)). [56]
MCS can be classified into Maximum Common Induced Subgraphs (MCIS) and Maximum Common Edge Subgraphs (MCES).

- Maximum Common Induced Subgraph MCIS
  If $G1$ is a subgraph of $G2$, then $G1$ is included in $G2$. The focus is on the maximum common induced subgraph. The common induced subgraph of $G1$ and $G2$ is $G12$. [55]
- Maximum Common Edge Subgraph MCES
  MCES is a subgraph that consists of the largest number of common edges of $G1$ and $G2$. [55]

Also the MCS can also be divided into connected and not connected MCS.

- connected MCS
  A connected MCS consists of a common allied subgraph. Each vertices is connected to every other vertices by at least one path in the graph.

Usually it is exhausting manual grouping compounds from large databases. Automatic cluster methods can be helpful, which have similarity metrics for the comparison of structures on the one hand and a cluster algorithm on the other, which sorts and classifies the compounds into structurally related groups. [55]

- disconnected MCS
  A separate MCS consists of two or more separate subgraphs. Disconnected MCES are mainly used to find similarities between compounds that would otherwise not have large common substructures, but rather look at common functional groups. [55]

The determination of the reaction center can be made possible by the MCS structure, since one can determine the bonds that are not contained in the MCS. During the reaction, these bonds are detected as broken or newly formed bonds. [56]

Extended connectivity maximum common substructure methods (EC-MCS) based method:

Lynch-Willett method: This method identifies reaction centres based on the EC value of the Morgan algorithm. First, it is iterated until the EC values for all educt and product atoms have no atom pairs (EC $^{n}_{ri}$= EC $^{n}_{pj}$). To obtain the EC values, the Morgan algorithm is used to identify equivalent atoms within a molecular structure. Each atom is first given an initial EC (EC^{0}_{i}) value. The EC value corresponds to the number of neighbouring non-hydrogen atoms. Then add up the various EC values and obtain the value for the number of equivalence classes k. This process is repeated until $k' \leq k$. Then the process is ended. [57]
The educt and product atom pairs are now labelled, which are located in the center of identical circular substructures with a radius of n-1 bonds (this is the EC-MCS). Then eliminate those atoms with the same EC^{ji-1} on the educt and the product. The steps are repeated until all atoms have been removed that were contained in an EC-MCS, after that the reaction center can be identified and atom mapping can be produced. [58]
The principle of the Lynch- Willett approaches is illustrated in 1.40:

[16] [52]

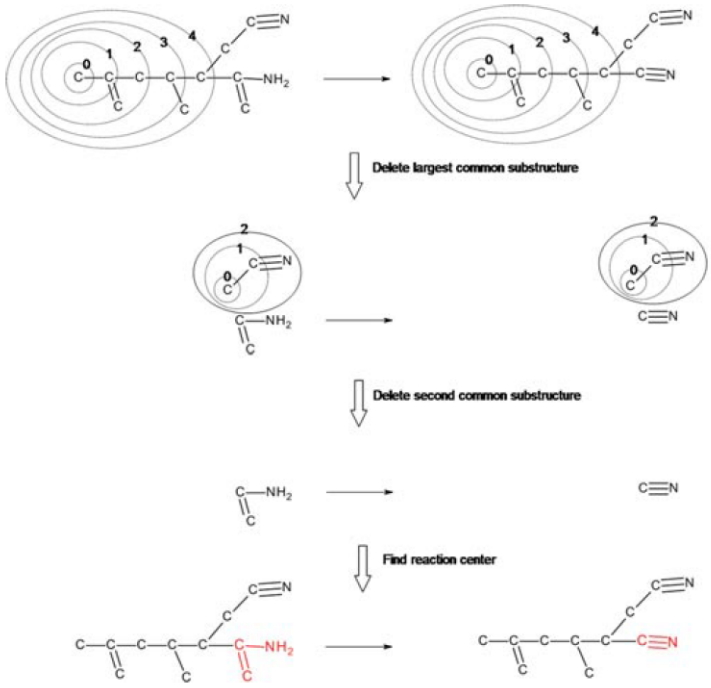


Figure 1.40: Lynch-Willett method: After the fourth iteration step, a large substructure is obtained which contains the radius of four bonds. Now this substructure is eliminated in both the educt and the product side. This process is repeated and after removing the CCN substructure, which consists of a radius of 2 bonds, the reaction center is obtained. [16]

Unfortunately, the method has its limits. For small molecules with large reaction centres, this method cannot be handled. Reactions that would lead to ambiguous atom mapping cannot be processed with this approach either.

Maximum common substructure (MCS) based methods:

McGregor–Willett's MCS-Based algorithm: This approach is based on MCS, which consists of two stages and should enable the identification of the reaction centres. First, the raw reaction center is searched for using the Lynch Willett method. Secondly, the MCS is then searched for in the reaction center using a modified MCS process. The MCS algorithm that is used is based on the backtracking method of McGregor. The backtracking method is a refinement of the brute force method. With the backtracking method, the MCS is determined from the two graphs. The algorithm finds solutions and discards a solution if the algorithm determines that this solution does not fit. Trees are used to represent the solutions. You start at the root node of the tree and select an option; if this option cannot be accepted, the algorithm returns to the last starting node and selects another edge. This is done within the tree until the solution is found. [59] A modified MCS procedure is used to facilitate the identification of the reaction

center. The MCS can consist of several fragments to facilitate identification of bonds that are broken or formed during the reaction, and two bonds of different order can match each other. Now you identify the changes in the binding order and the reaction center. Finally, you get the atom mapping. The efficiency of the MCS search is increased with this method, since the preliminary reaction center is much smaller than the total reacting molecules. The limits of this approach are in unbalanced chemical reactions and incorrect atom mapping can be generated for certain reactions such as hydrolysis and etherification. [60]

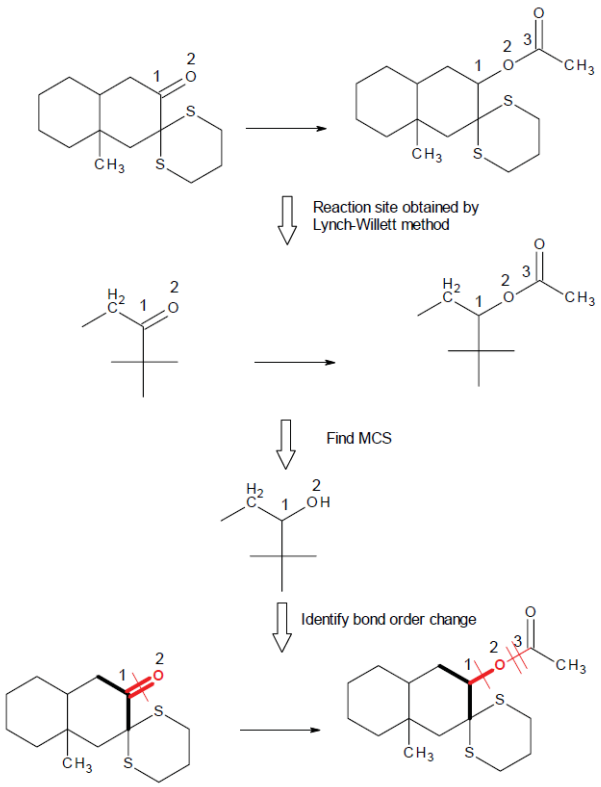


Figure 1.41: McGregor–Willett method: First, in step one, the Lynch Willett method is performed on the reaction in order to obtain the reaction center. In step two, the MCS is determined using the McGregor-MCS search method. In the third step, the MCS is compared to the reaction sites of the reactant and the to the product. Here you can see that atoms 1 and 2 in the product must come from atoms 1 and 2 of the educt. During the reaction, the double bond between atoms 1 and 2 of the educt was converted into a single bond between atoms 1 and 2 in the product. Since atom 3 of the product is not contained in the MCS, the bond between atoms 2 and 3 of the product must be re-formed in the course of the reaction. The reaction with the reaction center is shown in the last step.[17]

[17] [52]

Maximum common edge substructure (MCES) based methods:

Körner and Apostolakis' algorithm: The Körner and Apostolakis approach is based on three assumptions and is also known as the ITSE-based method. The first assumption is that the reaction mechanism converts the educts into the products with the lowest activation energy (= low temperature assumption). The second assumption is that the reaction has a single transition state. The last assumption is the additivity assumption, i.e. the activation energy for the transition state (referred to as Imaginary Transition State Energy, ITSE) is the sum of the activation energies of reacting bonds. [61] With the help of the MCS of weighted edge graphs, the problem of minimizing the ITSE can be solved. The edge graph $E(G)$ of an undirected graph G is another graph $E(G)$, which represents the neighbouring edges of G (= line graph). With the line graph, the edges become the nodes. Thus, the nodes of a boundary graph represent the bonds of the original molecular structure. The Raymond et al. was expanded by Körner and Apostolakis. Now the MCS-based algorithms have different weights. Similar to the previous MCS-based algorithms, they also use different weights for matching bonds with different multiplicity. Furthermore, weights are also determined depending on the type of atoms that form the bonds. CC sigma bonds weigh 1.5; CN amine, CO ester and CS thioester bonds have a weight of 0.48 and the remaining bonds have a weight of 1. [61]

1.5.3 Optimization-based methods

In the case of a chemical reaction, i.e. the transformation from educt to product, the shortest path is followed. [62] So atom mapping solution should be found with a minimum number of broken and formed bonds.

Heinone et al. algorithm: The A * algorithm is used when finding paths and traversing graphs. The algorithm searches for the best path using the best first search, which should have the cheapest path from a start node to a goal node. [63] The goal of the algorithm is to find an atom map that minimizes the graph edge edit distance. The edge edit distance is the minimum number of edge edit operations that have to be done to transform the first graph into the second graph. So in atom mapping, the edge edit distance is the minimum number of broken, newly formed and rearranged bonds in order to get the educt into the product during the reaction. The three most important components from Heinone et al. Algorithm are: An A * type total path cost estimate to guide the search in the space of partial atom mappings. An extension operator for partial mappings that maintains the path cost estimates in constant time per edge. Pruning of A * search space by computing upper bounds on the optimal cost via fast greedy search. [63]

$$f(n)=g(n) * h(n) \tag{1.5}$$

The additive evaluation function $f(n)$ estimates how long the best path is from the starting node to the destination node using the node under consideration. Where $g(n)$ describes the costs for the transition from the starting node to the

next node n. $h(n)$ describes the heuristic estimation costs for the transition from n to any other node. [64] [63]
By using the breadth first search the atoms of the reactants are numbered. The search begins with an external atom of the largest reactant. The remaining reactants are processed iteratively in the order of their size. This enables the determination of the minimum number of broken and formed bonds that is needed to transform the educt into a product. However, there is also limitations that all reactions have to be chemically balanced. [65]

Linear programming (LP) or linear optimization is a mathematical method in which the best way is determined so that the best result, i.e. the maximization or minimization of a linear objective function, is achieved in a mathematical model that is subject to a group of linear equation constrains. With the LP method, a point is found in the polyhedron at which the objective function receives the best (e.g. smallest or largest) value (if this point also exists). [66]
Integer Programming (IP) or integer linear programming problem is when all unknown variables must be integers. It is referred to as a MILP problem when only some of the unknown variables have to be integers. Both are NP-hard, which means that it is not possible to get a deterministic solution in the polynomial time.
A MILP formulation with binary (0,1) variables [67]:

$$\begin{aligned} min\ \ & c^T x + d^T y \\ s.t.\ \ & Ax + Bx \leq b \\ x \geq 0,\ \ & x \in X \subseteq \mathbb{R}^n \\ & y \in 0, 1^q \end{aligned}$$

x ... vector of n continuous variables
y ... vector of q binary (0,1) variables
c, d ... vectors of parameters
A, B ... parameter matrices of appropriate dimension
b ... vector of p inequalities

The conditions of the matrices A and B must be fulfilled by the vectors x and y. The basis for maximizing the objective function are the vectors of the parameters c^T and d^T, which represent transposed vectors. The matrix A represents all atoms that are involved in the reaction. The matrix B represents sets of all educt and product bonds. The number of atoms in the educts and products should be the same and each atom should only be mapped with the same atom type. Each bond is given a coefficient that represents the probability of a bond breaking or bond formation, so the objective function can use these coefficients to determine the minimum costs for bond breaking and bond formation. It moves on to the integer linear programming problem LP if vector c and matrix A or if vector d and matrix B have no elements. [67]

First et al Algorithm: The reaction mapping problem is expressed as a MILP model. This approach minimizes the number of bonds broken, bonds formed, and changes in the bond order between substrates and products. The objective

function is composed of four summation terms, the first summation term being for the reaction bonds. Each term is equal to one when the bond is broken. The second term for the product bonds is also equal one when a new bond is formed. The third term is about the tetradic atoms, here the term assumes equal one when the stereochemistry changes. The fourth term is for the stereochemical double bonds, here the term assumes equal one if the stereochemistry changes. The objective function can be viewed as the total number of bonds that break and form. There are eight restrictions on the MILP model. With restriction 1 and 2, there must be a one-to-one mapping of the atoms in the reactants to the atoms in the products. With restriction 3, atoms of the same type must be mapped onto one another. With restriction 4 and 5, a variable with the value one is defined when the reactant binding is assigned to a product binding. Changes in stereochemistry during the reaction are defined by constraints 6, 7 and 8. Solutions of the model are obtained in which a one-to-one mapping between educt and product atoms is obtained. First's algorithm is able to find several optimal maps. Equivalent maps can be traced back to symmetries of the reaction molecules. To solve the problem, a technique is used that numbers both pairs of equivalent atoms and maps them with an index. Here the equivalent atoms are shown in pairs with the same index. [52]

Latendresse et al. Algorithm: MWED uses also a MILP approach that aims to minimize changes to bindings in biochemical reactions. The algorithm is a combination of First et al. algorithm with bond weights. Weights are assigned to the bindings and specific cost when a binding is modified. For instance, the weights for a C-C bond are 4, for a C-N bond 0.56, for a C-O bond 0.48. The lower the weight, the easier it is to break or create the bond. Several optimal mappings can be generated, but the atom mapping that has the minimal MWED is selected. The algorithm uses bonds with hydrogen atoms and steroid chemistry to generate the atom maps. Chemical equivalent atoms and reaction centres are identified. Hydrogen atoms are not formed, but are taken into account when calculating the atom maps. RXN and SMILES can be used as input files, which create SMILES and MetaCyc output files. [15] [68]

1.5.4 Atom mapping tools

In the following section five atom map tools are presented, which identify the atom maps of reactions on different prediction paths respectively by using different algorithm.

AutoMapper: One approach that uses the maximum common substructure (MCS) and minimum chemical distance (MCD) to predict atom mapping is the AutoMapper. Here the largest substructures of the substrate graph are found, which are isomorphic to the product graph. Those atom maps of the atoms that are not part of the isomorphic substructure are calculated by MCD, that minimizes the number of bonds that are broken and formed. Automapper also takes into account the presence of alternative substructures or products in the case of reactions that are not chemically balanced. Another advantages of Automapper is that it can also handle stereochemistry. Automapper also displays hydrogen atoms maps, but cannot identify chemically equivalent atoms or reaction centers. Many formats can be used as the input file. For example RXN,

InChI and SMILES. The generated atom mapping is displayed as an output file in RXN or SMILES format. [52] [68]

DREAM: A web tool that identifies atom mapping using the optimization-based approach of MILP is Determination of REAction Mechanisms (DREAM), which is freely available. The atom maps are determined from RXN or SMILES files, which also depict hydrogen atom maps. The reaction must be chemical balanced for DREAM to accept these reactions. The input file can either be an RXN file, a SMILES or it is also possible to draw reactions with the interactive editor. The output file is a RXN file. The RXN file don't have any information about the reaction center, but atom mapping information. After the mapping is performed, the results are sent back to the e-mail. [68]

RDT: Reaction Decoder Tool (RDT) is an open-source java based atom mapping tool, which returns the four best atom mappings, which were created according to four different algorithms. The algorithms would be Mixture MCS. This corresponds to the maximum common substructure matching between educt and product; Min-Sub-model, which corresponds to the smallest substructure matching between educt and product; Max-Sub-model, which corresponds to the largest substructure matching between educt and product and the assimilation model is used if a ring system is present. It is possible to map both chemically balanced and chemically unbalanced reactions. After the algorithm has reached a maximum number of atoms, the remaining atoms are mapped with a similarity rating and the selection and elimination steps is repeated until all atoms are mapped. This model provides valid atom mappings that create a minimal number of binding changes. It is a combination of graph theory and mathematical optimization of algorithmic results that are generated by a series of chemical rules. It is possible to determine reaction centres with RDT, but hydrogen atoms and equivalent atoms cannot be recorded. The input and output files are RXN or SMILES. [69] [70]

ICMAP: InfoChem-Map (ICMAP) is a software tool that can identify reaction centres using MCS and MILP approaches. The software recognizes reaction centres by using minimal chemical distances when bonds are broken or re-formed. To support the MILP approach for finding the best possible mappings, some additional chemical rules are applied. When breaking and forming bonds between heteroatoms, preference is given to C-C bonds and bonds with hydrogen atoms are evaluated in the same way as C-C bonds. ICMAP has limitations in terms of the number of molecules in the reaction, the size of the molecules and the single atom mapping is limited. In addition, ICMAP cannot map atoms in which all chemical bonds are broken and newly formed. Chemical reaction centers are obtained, but you cannot determine equivalent atoms or determine hydrogen atoms. The input and output files are in RD format. [71] [68]

CLCA: Canonical Labelling for Clique Approximation (CLCA) can determine the maximum common structure between educt and product by using prime factorization. This creates canonical markings for bond atoms. MCD is used to select the substructure that reduces the number of bond changes between educts and products. Because there are many combinations of MCS when a reaction has many reactants or product graphs. Canonical markings are formed

using chemical properties. CLCA can determine chemical equivalent atoms and reaction centers. SMILES is used as input and output files. [72]

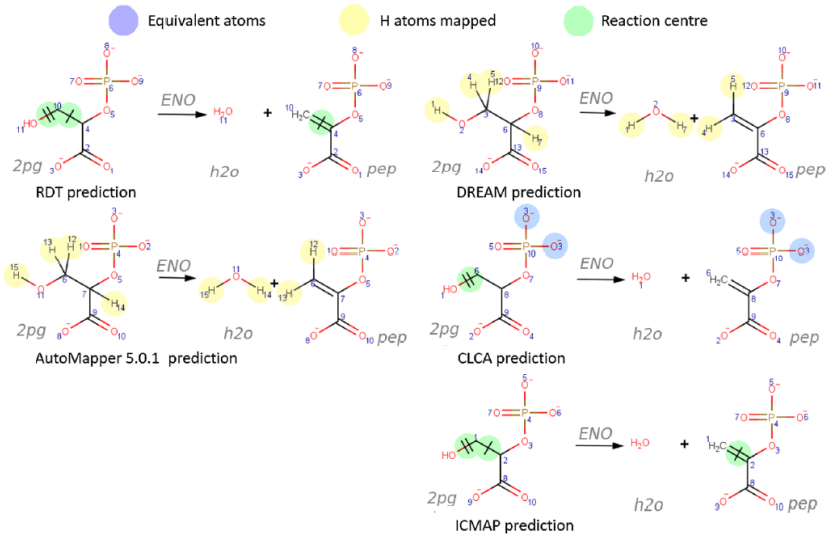


Figure 1.42: In this figure the predictions of different atom mapping tools of the enolase reaction are presented. Equivalent atoms are labelled with a blue circle. Hydrogen maps are marked with a yellow and reaction centres are marked with a green circle. All tools, which are presented, create atom mappings of the reaction. In addition to atom maps, the tools can generate additional information such as hydrogen mapping using DREAM and AutoMapper; reaction centers can also be created with RDT, CLCA and ICMAP or atoms equivalent to CLCA.[18]

[18] modified: [68]

1.6 Similarity search

There are several approaches to investigate the similarity of the molecule, the choice of method depends on the property of the molecule, which want to be compared. Usually, the structural similarity of molecules is of great interest to chemists. Descriptors for molecules are used to enable a fast search for similarity in large databases. Descriptor fingerprints can be pre-calculated and stored when a molecule is added to the database. So the fingerprints can be used to find similar molecules from the database. The fingerprints are described in the section 4.1. The similarity between graphs can also be measured using the maximum common subgraph approach, which is described in more detail in section 1.5.2.

1.6.1 Fingerprint

The fingerprints encode the molecular structure of a chemical species in a sequence of binary digits that represent the presence and absence of a substructure in a chemical molecule (the fingerprints contain the structural information about the molecule). An advantage is that 2D/ 3D features of a molecule is encoded as a set of binary values, which is computational more efficient. With fingerprints a similarity search between two molecules can be done by comparing these fingerprints with each other.

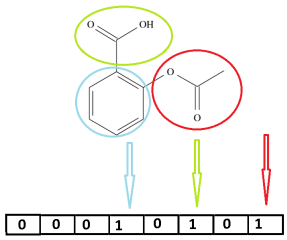


Figure 1.43: Fragment Code Fingerprint

Tanimoto coefficient

A common way to calculate the similarity is with the distance measure Tanimoto coefficient. It is a measure of the number of common substructures shared by two molecules.

$$\mathrm{T}(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{1.6}$$

The fingerprint is used to explain the Tanimoto coefficient. In the similarity measure Tanimoto A and B are sets of fingerprints of molecule A and molecule B. $|A \cap B|$ is the set of common fingerprints of both molecule A and B. $|A \cup B|$ is the union of fingerprints of both molecules A and B. [73]

The Tanimoto coefficient ranges from 0 to 1, where 0 means no similarity between the two molecules/ fingerprints and 1 means that the molecules/ fingerprints are identical. [74]

Chapter 2

Methods

2.1 From atom mapping SMILES to graph rewrite rule

To have the biochemistry fully accessible as a graph rewrite rule, the first step was to get the atom mappings of the biochemical reactions. All atom mapping reaction SMILES were downloaded from the *MetaCyc* database. These atom mappings from *MetaCyc* are stored in one flat file named atom-mappings.dat. These atom maps were determined from *MetaCyc* by using the MWED method within Pathway Tools.

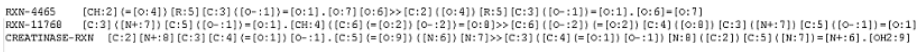


Figure 2.1: That downloaded file contains one reaction per line, the first element on each line is the frame ID of the reaction, followed by a tab, then one atom mapping reaction SMILES separated by a space. A total of 14,000 atom mapping reaction SMILES are included in the downloaded flat file.

In the section SMILES the syntax of the SMILES string is described. However, there were some SMILES strings in the downloaded file that do not follow the SMILES standard. Therefore the invalid atom mapping reaction SMILES had to be first fixed before parsing it to the rewrite rules.

1. One issue is that charges in the reaction SMILES should be written with a number instead of multiple "-" or "+". For instance, "[Fe+]", "[Fe+2]", and "[Fe+3]" are valid SMILES strings for describing charges, but "[Fe++]" and "[Fe+++]" is not in the grammar.
   - The invalid multiple charges of the atoms are changed to the valid SMILES syntax, so the charge chains of "+" and "-" were conformally converted into signs followed by numbers.
2. Another big issue is that the downloaded atom mapping reaction SMILES use abstract atoms respectively whole names of functional groups or proteins, which are invalid for the SMILES syntax.
   - All invalid or "strange" atomic labels and whole names of functional groups have been converted to wildcard "*".
4. Sometimes more than one reaction SMILES appears separately for a the same reaction name.
   - So, a separate line and a postfix (i.e., _a, _b, _c, etc.) to the reaction name are added.
5. On of the biggest problems was that the oxygen atoms in the acids- and phosphate- groups of the reaction SMILES were labelled with the same IDs.
   - In order to obtain different numbering, the acid groups had to be matched with SMARTS and given internal IDs.
6. The downloaded atom mapping reaction SMILES have also a lot of reactions, which aren't chemical balanced, there are many missing atoms and molecules.

  - This problem is still open and needs to be resolved for all unbalanced reaction SMILES. For the later construction of the network for the biosynthesis of the 3-hydroxypropanoate (3HP) from pyruvate, the selected rewrite rules were manually improved so that they are chemically balanced.

7. In addition, the educt and product sides of the atom mapping reaction SMILES are reversed in some cases. That means on the left side where the educts are usually given, sometimes products are added and on the right side where the products are usually placed, educts are for some cases added.

   Furthermore, a *RDKit* script was created, which has a SMILES parser that is more general than in the network construction tool MØD that accepts some that MØD would not take. This *RDKit* script parsed into canonized SMILES. Next the canonized SMILES were parsed by MØD again in SMILES, so to get SMILES that can read by MØD in any case and convert them back into graph rewrite rules.

8. With 25 reaction SMILES (RXN-8787, 1.8.5.2-RXN, RXN-7673, RXN-8790, RXN-7677, RXN-8793, 1.17.5.1-RXN, RXN-12581, RXN-8786, RXN-8789, SHIKIMATE -PQQ-RXN, RXN-7676, RXN-8792_a, RXN-8792_b, 1.10.99.2-RXN, RXN-8785, RXN-8788, 1.1.99.25-RXN, RXN-8791, RXN-13352, RXN-13351, RXN-13358, RXN-13361, TRANS-RXN0-488, RXN0-1702) an error appeared when parsing with a *RDKit* script, namely that the valence of a C, N or O atom was too large. These are all reaction which have either metal complexes or "strange" additional protein binding sites on O, C or N atoms.

   - These listed reaction SMILES has been removed and left out.

After fixing most of the described issues the valid reaction SMILES were parsed into the maximum graph rewrite rule. The maximum graph rewrite rule contains all atoms and bonds, which are involved in the atom mapping reaction SMILES. This rule is the most specific rule, because it involves the complete environment of the atoms.
Then the minimum rules were created from the maximum rules, these minimum rewrite rule only contains those atoms on which the bonds are broken or formed directly, so it is a rule of the reaction center.

2.2 Database

Data used in ERRD is available as a SQLite file and the image 2.2 shows the associated SQL schema.

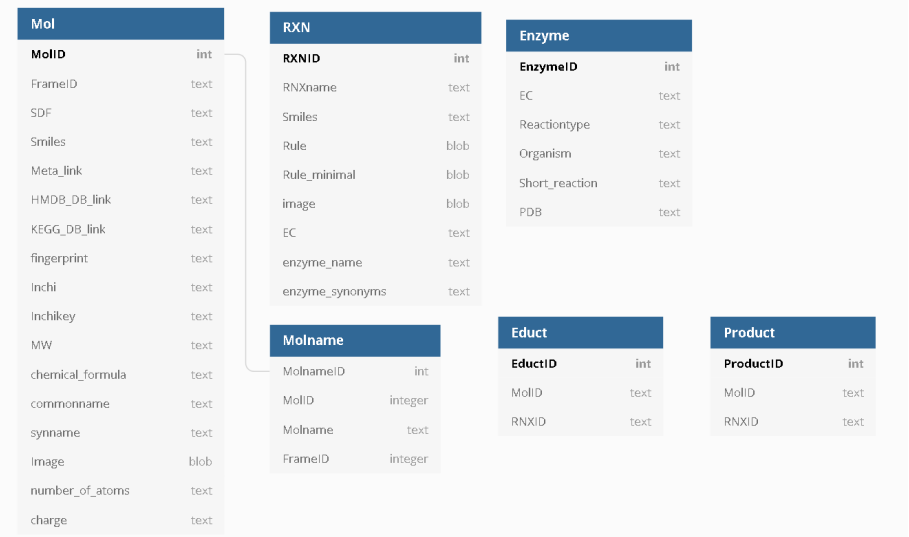


Figure 2.2: Six tables have been created with the names *Mol, RXN, Molname, Enzyme, Educt and Product.* The *Mol* table begins with the primary key MolID, followed by the text-based entries that contain useful information about the molecule. For example, various line notation formats such as SMILES, InChI and InChIKey are entered, but also the connectivity table SDF. In addition, various molecular properties such as the molecular weight, the chemical formula, the common name, synonyms of the compound, the number of atoms and the charge have been entered in the *Mol* table. The binary images that represent the molecule on the website were saved as a blob file. In addition, various external links of the compound are entered in the *Mol* table. The information was extracted from various databases.
The *RXN* table begins with the primary key RXNID, followed by the RXN name. In the *RXN* table the atom mapping reaction SMILES is entered as a text file and both the minimum and the maximum graph rewrite rule of the reaction is entered as a blob. The binary image of the enzymatic reaction is also saved as a blob in the *RXN* table. The EC number of the reaction, the enzymatic reaction name and the associated synonyms are also entered in the *RXN* table.
The *Enzyme* table contains the EnzymeID as a primary key, followed by information about the enzyme as a text-based file. Such as the EC number, the type of reaction, the organism in which the enzyme is present, and the reaction equation consisting of the compound names and the PDB file.
The tables with the names *Educt* and *Product* both have the EductID or the ProductID as the primary key. Both tables contain the MolID and RXNID if they occur as a educt or product.
The *Molname* table starts with it's primary key MolnameID, followed by the MolID and FrameID as integers. Within this table different names or synonyms are stored for each compound

2.3 Technology

The ERRD website has been implemented using Flask Python (https://flask.palletsprojects.com) as Web Framework with a SQlite database (https://www.sqlite.org). Jinja (https://jinja.palletsprojects.com) was used as Flask's standard template engine. The database schema is shown in an UML diagram. Python (https://www.python.org/) is the language which is used to write the scripts. JavaScript and CSS are used on the client side. The two dimensional images of the molecule are created by *ObenBabel* (http://openbabel.org/docs/dev/OpenBabel.pdf). The two dimensional image of the biochemical reaction with the atom mapping annotation was created by using the web tool *CDK depict* (https://www.simolecule.com/cdkdepict/depict.html), which creates depictions of molecules and reactions by taking the atom mapping reaction SMILES as input. The three-dimensional images of the molecules and enzymes on the website are created by 3DMol.js (https://3dmol.csb.pitt.edu/). The chemical connection table files (SDF, PDB and TPL) and the SMARTS are created by *RDKit* (https://www.rdkit.org/). The molecule information (name, synonyms, formula, molecular weight, charge, canonical SMILES, InChI, InChIKey and fingerprint) are filtered and created by *OpenBabel* from the SDF files of each compound. The listing of the compounds consumed as educts respectively produced as products during a reaction is identified and found using the method of the common subgraph isomorphism.
To generate the graph rewrite rules from the atom mapping reaction SMILES and to build in the next step networks, the generative network construction tool MØD (https://jakobandersen.github.io/mod/) was utilized for that purpose. The atom mapping reaction SMILES were downloaded from the database *MetaCyc*.

Chapter 3

Results and Discussion

3.1 Website

Figure 3.1: ERRD Logo: In this picture the logo of the database is shown. The abbreviations stand for enzymatic reactions rule database.

In the following chapter the website of the ERRD and its possibilities and functions are described.

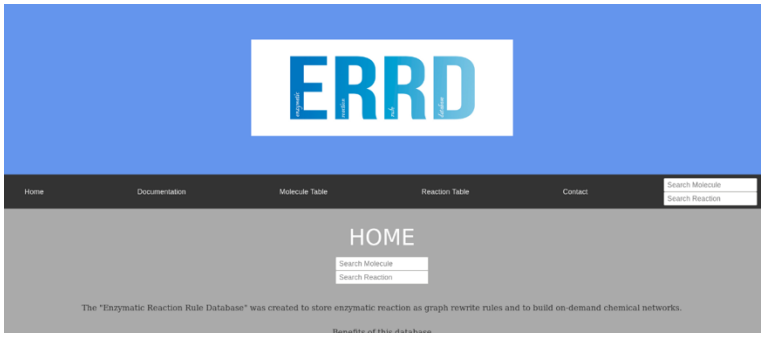


Figure 3.2: The user can search for the molecules or the enzymatic reaction on the search bar by entering either the molecule ID or the molecule name respectively the reaction ID or the reaction name. The other pages, which are described in more detail in the next sections, can be accessed through the navigation bar.

3.1.1 Molecule table

On the molecule table page there is a table which lists all the molecules that are stored in the database. The line number is shown in the table first. Followed by the molecule ID column and the column with the name of the molecule. Furthermore there is the column with the respective SDF format, which can be downloaded directly from the website. The last three columns are the links that lead to the MetaCyc, HMBD and KEGG websites. If a link is present, it is marked with "Link", if a link for the respective compound is not available, it is labelled with "No Link". An extra window of the linked website opens when the user click on the website links. It is also possible to change the row numbers of the tables to 10, 25, 50 or 100 in order to get an expanded overview.

Figure 3.3: The complete table of the molecule table page is illustrated. The line number and the molecule ID, which leads to the respective molecule information page, are listed. The name of the molecule is given in the next column. It is also possible to download the SDF of the respective molecules. This table can also be used to link to various databases (for example MetaCyc, KEGG and HMDB).

With the search bar of the molecule table the user can search for the molecules either by molecule ID or by molecule name. Only those molecules are listed that have the complete name or a part of the name. This saves time and is a simple method to find the desired molecule as quickly as possible.

Figure 3.4: The search function is shown in this illustration. The molecule can either be searched for using the molecule ID, as shown in the picture on the left, or by molecule name, as shown in the picture on the right.

Above the table is a download button with where the user can download all SDFs of all molecules in one file. By clicking on the molecule ID the user will be taken to the information page of the molecule.

Figure 3.5: SDF file from all molecules can be downloaded from the ERRD.

3.1.2 Reaction table

All enzymatic reactions that are stored in the ERRD database are listed on the reaction table page. In this table, the line number follows first, followed by the reaction ID, then the column with the name of the enzymatic reaction, followed by the two download buttons of the maximum and minimum graph rewrite rule of the reaction.

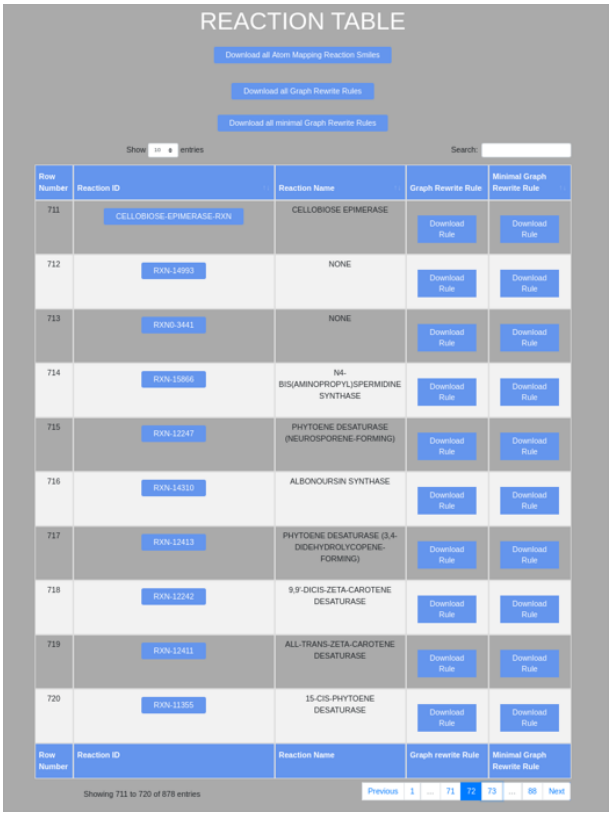


Figure 3.6: The complete table of the reaction table page is shown. The line number and the reaction ID that leads to the reaction information page are given. Followed by the reaction name and the minimum and maximum graph rewrite rules. Both of these graph rewrite rules can be downloaded.

Above the reaction table, the user can directly download a single file with all atom mapping reaction SMILES, also a file with all maximum graph rewrite rules or a file with all minimum graph rewrite rules. Just like the table in which all molecules are listed, the number of reactions shown in the reaction table can also be expanded by 10, 25, 50 or 100 items.

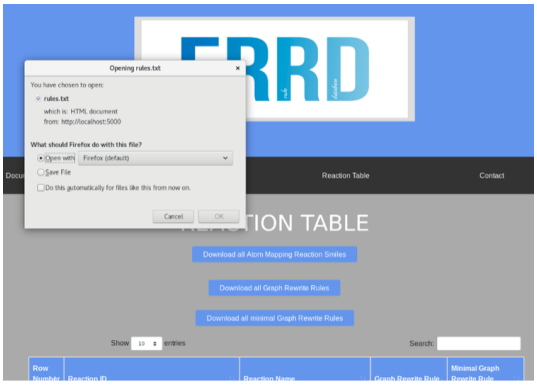


Figure 3.7: SDF, maximum and minimum graph rewrite rules files form all enzymatic reactions can be downloaded from the ERRD

Within the table a search bar has been inserted that should make it easier to find reactions in the table. With the search bar of the reaction table the user can search for reaction either by reaction ID or by reaction name. By clicking on the reaction ID button the user will be taken to the information page of the reaction, which is explained later in detail in this work.

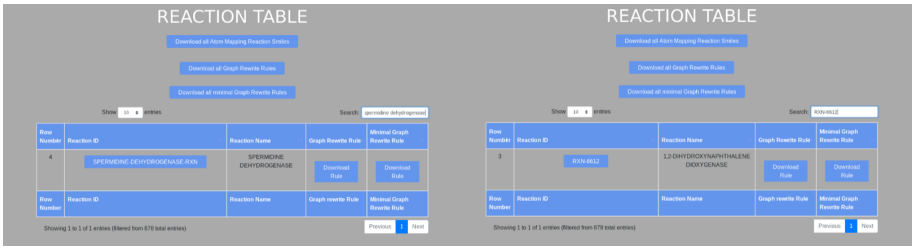


Figure 3.8: The search function is shown in this illustration. The enzymatic reaction can either be searched for using the reaction ID, as shown in the picture on the left, or by reaction name, as shown in the picture on the right.

3.1.3 Information of the molecule

The user can get the general information of a molecule in the *information page* of the desired compound. Such as the molecule ID, the name, the synonyms, the chemical formula, the molecular weight and the charge of the molecule. So the user can get a brief overview of the molecule. The information is generated by OpenBabel using the SDF of the molecules as input file.

Molecule ID	CPD-6993
Name	pinocembrin
Synonyms	None
Formula	C15H12O4
Molecular Weight	256.253 Daltons
Charge on the molecule	0

Figure 3.9: The table shows general information about the molecule of interest. Information such as the molecule id, the molecule name, the known synonyms, the chemical formula, the molecular weight and the charge of the molecule are listed. This example shows pinocembrin with the molecule ID CPD-6993. With the chemical formula $C_{15}H_{12}O_4$, a molecular weight of 256.253 Daltons and a non-existent charge, which is specified as zero.

On the right, next to the general information table of the molecule, the user can see a two-dimensional representation of the molecule, which is labelled with the molecule ID. The representation can be downloaded as a PNG file directly from the website. In addition to the two-dimensional image, the user can consider a three-dimensional representation of the molecule. This was created using 3DMol.js by implementing to the ERRD website. The representation of the molecule is in the drawing method "Stick". The molecule can be rotated by holding the mouse button on the surface and moving it in one direction. The molecule can be enlarged or reduced by rotating the mouse scroll wheel. It is also possible to translate the image by holding down the middle mouse button and moving the mouse in the respective directions.

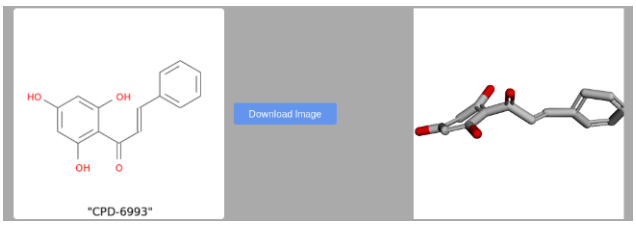


Figure 3.10: On the left side the two-dimensional image of pinocembrin with the molecule ID CPD-6993 is shown and on the right side the interactive three-dimensional image with the drawing method "Stick" is presented.

Under the two images of the molecule, the user can now download some text-based formats directly from the ERRD website. For instance, 2D SDF, 3D SDF, PDB and TPL. The formats, which can be downloaded, were created with the chemical tool RDKIT. These text-based formats can be very useful for further work as they contain coordinates of the atoms and the bonds among the atoms.

Figure 3.11: It is possible to download useful text-based formats from ERRD. The user can download SDF, PDB and the TPL format of the molecule.

In addition, the user will also find the canonical SMILES line notation, which is created with the help of OpenBabel, and the substructural patterns language SMARTS, which was created with RDKIT. These chemical representations of molecules can also be useful for further work or for chemical and biological projects. The fingerprint, the InChI and InChIKey formats, which were also created with OpenBabel, are also accessible.

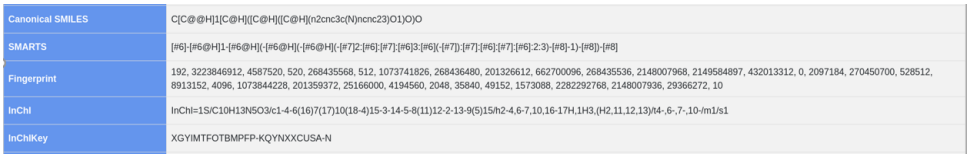


Figure 3.12: Further helpful information of the desired compound can be found in the information page. For example canonical SMILES, SMARTS, Fingerprints, InChI and InChIKey.

If the user wants to receive or collect more information about the molecule, it is also possible to be lead to other websites by clicking on the following links. For instance MetaCyc, HMDB or KEGG.

Figure 3.13: By clicking on the links, the user will be taken to the other well-known websites. For example MetaCyc, KEGG and HMDB.

Then there is also a listing of the reactions in the ERRD website, which on the one hand indicate in which reactions the molecule is consumed as a educt and on the other hand in which reaction the molecule is produced as a product. The RXN IDs of the reactions are given in the list. If the user clicks on these listed reaction links of the educts and products, they will be lead to the respective information pages of the reactions. There the user can now get further useful information about the reaction, which is explained in more detail in the following paragraph.

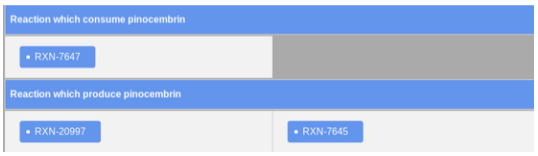


Figure 3.14: The listing of reaction, which consume and produce the compound pinocembrin is shown in this picture. In this case, pinocembrin is consumed through the reaction with the ID RXN-7647. The reactions with the ID RXN-20997 and RXN-7645 produce the compound pinocembrin during the chemical reaction.

3.1.4 Information of the reaction

The very first thing the user can see on the information page is the name of the reaction as a heading. The reaction equation is written under the heading in the form of the reaction names. The two-dimensional image of the biochemical reaction with the atom mapping annotation, which can be seen on the ERRD website, was created by using the web tool *CDK depict*, which creates depictions of molecules and reactions by taking the atom mapping reaction SMILES as input. The picture of the reaction can also be downloaded directly from the website as a PNG file by clicking on the download button.

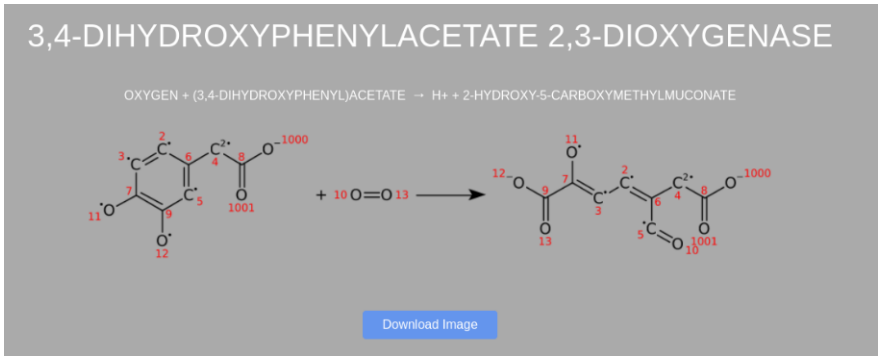


Figure 3.15: The two-dimensional image of the enzymatic reaction 3,4 dihydroxyphenylacetate 2,3 dioxygenase is shown. The reaction is labelled with their atom mapping annotation, so it is possible to trace individual atoms through the biochemical reaction. So, with the web tool CDK depict two-dimensional depictions of reactions by using the atom mapping SMILES as input, which can have carbon atoms with adjacent bonds, but some missing hydrogens, can be created. The missing atoms works fine for reaction rules pattern, but if CDK creates those images, black dots are created in the reaction, which represent the missing atoms, which are often hydrogen atoms.

For very few reactions, the enzyme that is involved in the reaction is also shown in a three-dimensional illustration. The three-dimensional image of the enzyme was created with 3Dmol.js. The representation of the enzyme is in the drawing method "Stick". The enzyme can be rotated by holding the mouse button on the surface and moving it in one direction. The enzyme can be enlarged or reduced by rotating the mouse scroll wheel. It is also possible to translate the image by holding down the middle mouse button and moving the mouse in the respective directions.

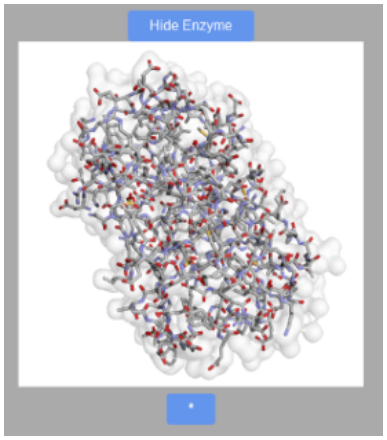


Figure 3.16: The three-dimensional image of the enzyme D-Lyxose Ketol-Isomerase is illustrated.

The unique ID of the reaction and the atom mapping reaction SMILES can also be found on the website. The graph rewrite rule, which was created using the chemical tool MØD, can be downloaded directly from the website. This applies to both the maximum and the minimum rewrite rule.

Figure 3.17: The information of D-Lyxose Ketol-Isomerase reaction are listed in this row. For instance, reaction ID, atom mapping reaction SMILES, download button for minimum, maximum graph rewrite rule and SMIRKS is given.

Some information of the enzyme involved in the reaction are also shown on the website. The EC classification number is given. In addition, the link to the well-known BRENDA database can be accessed in order to obtain further information about the enzyme. The enzyme name, the synonyms, the reaction type and in which organisms the enzyme is located can also be found in the table of the ERRD website. If no information is available, "None" appears in the line.

Figure 3.18: The information of the enzyme of the reaction D-Lyxose Ketol-Isomerase are listed in this row. The EC number, the link to BRENDA, the enzymatic name, the synonyms of the enzymatic reaction, the type of reaction and the organism in which the enzyme is present can be viewed.

The educts and the products involved in the enzymatic reaction have been listed.

The reactants are identified by the molecule ID of the respective molecules. Clicking the molecule ID button takes you to the information page of the molecule, which was previously described in detail.

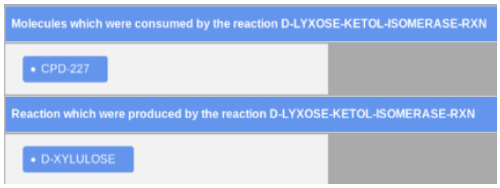


Figure 3.19: List of the molecule ID of the educts and products, which were consumed respectively produced by the reaction of D-Lyxose Ketol-Isomerase. The enzymatic reaction D-Lyxose Ketol-Isomerase consumes the molecule with the ID CPD-227 and produces the molecule with the ID D-XYLULOSE.

3.2 Possible applications and conclusion

A possible application would be that maximum and minimum rewrite rules can be taken from the created database ERRD and fed into MØD to generate networks, which can be used to find known and novel pathways for the biosynthesis of valuable organic chemicals.

With the construction of complex reaction networks, the research of promising biochemical alternatives, which were catalyzed by known enzymes, can be investigated. This approach leads to known and novel production of compounds of interest. [75]
The design of biosynthetic alternatives are getting more interesting, because the biochemical synthesis have more advantages over the organic synthesis. For instance, biochemical synthesis are more sustainable than the organic synthesis, because renewable feed stock may be used under mild temperature and pressure conditions. Also other advantages are a high yield of the compound and less by-formation production. [75] [76] [77] [78]
The biochemical production of 3-hydroxypropanoate (3HP) from pyruvate is of great importance. 3-hydroxypropanoate is also used as a building block molecule, which is utilized as a raw material to create other high valuable chemicals and it is also used in numerous other applications. [79] [75] For instance, polyhydroxypropanoate can be created using the 3HP monomer or acrylic acid, 1,3-propanediol or acrylamide can also be produced from 3HP as raw material. [79] [80]

Brunk et al. [75] created the Biochemical Network Integrated Computational Explorer (BNICE) framework, which was developed to find and evaluate novel and already proposed biosynthetic pathways for the production of 3HP from pyruvate. A set of reaction rules and chemical species are given to generate the all proposed and novel pathways. 86 reaction rules are utilized, which are based on third level EC classes represented in KEGG. So, the reaction rules are generalized, to make them less specific and to be able to find novel alternatives. A network was generated, where the chemical species represent the nodes and the directed edges represent the reactions. Next, a depth first search was performed, which started from pyruvate, and reported new routes once the node corresponding to 3HP was discovered. 17 of the 86 reaction rules were required to generate all proposed pathways to produce 3HP from pyruvate and 5 additional reaction rules were required to produce the novel alternative routes. [75]

In the following figure 3.20 all proposed pathways for the biosynthetic production of 3HP from pyruvate where shown:

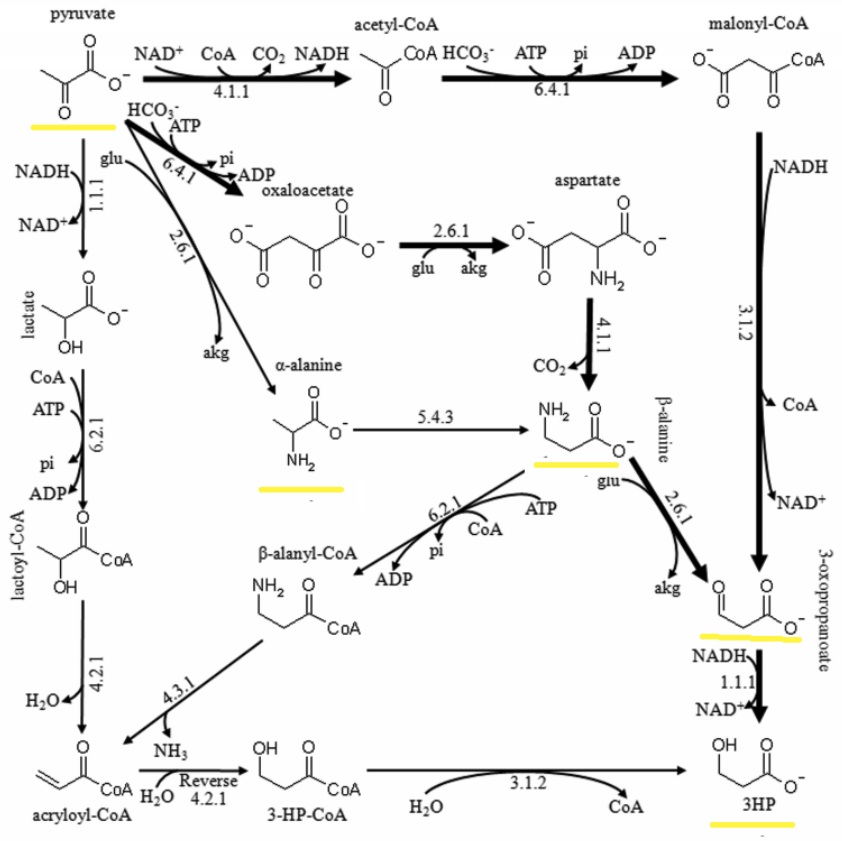


Figure 3.20: All known pathways for the biochemical production of 3HP from pyruvate are illustrated. The thermodynamic favourable are labelled in bold. The industrial implemented one is the four- step pathway with the intermediates pyruvate $\rightarrow$ α-alanine $\rightarrow$ β-alanine $\rightarrow$ 3-oxopropanoate $\rightarrow$ 3HP. Although some of the other pathways are thermodynamic more favoured, they aren't used in the commercial, because of the ATP consumption in one of their reaction steps, which decreases the maximum yield of 3HP that make these routes less economical. [1]

With the thermodynamic metabolic flux analysis the maximum yield and concentration of 3HP can be determined by using glucose as raw material. All reaction pathways, except the industrial implemented one, have at least one reaction step, which involves the consumption of ATP, which decreases the maximum yield of 3HP from glucose. And with each other mol ATP the yield of 3HP decreases significantly. So, the commercial pathway has the most advantages over all other known routes [75]

[1] [75]

To illustrate the possible application, 17 of the generated maximum rewrite rules were taken as input-rules from the created database ERRD and were utilized to generate the network of the proposed pathways of the biosynthetic production of 3HP by using MØD. All compounds and cofactors, which were included in the reaction rule were set as input-graphs. One iteration step were set to create the network. All proposed pathways could be reproduced by the maximum rewrite rules, which is shown in the figure 3.21.

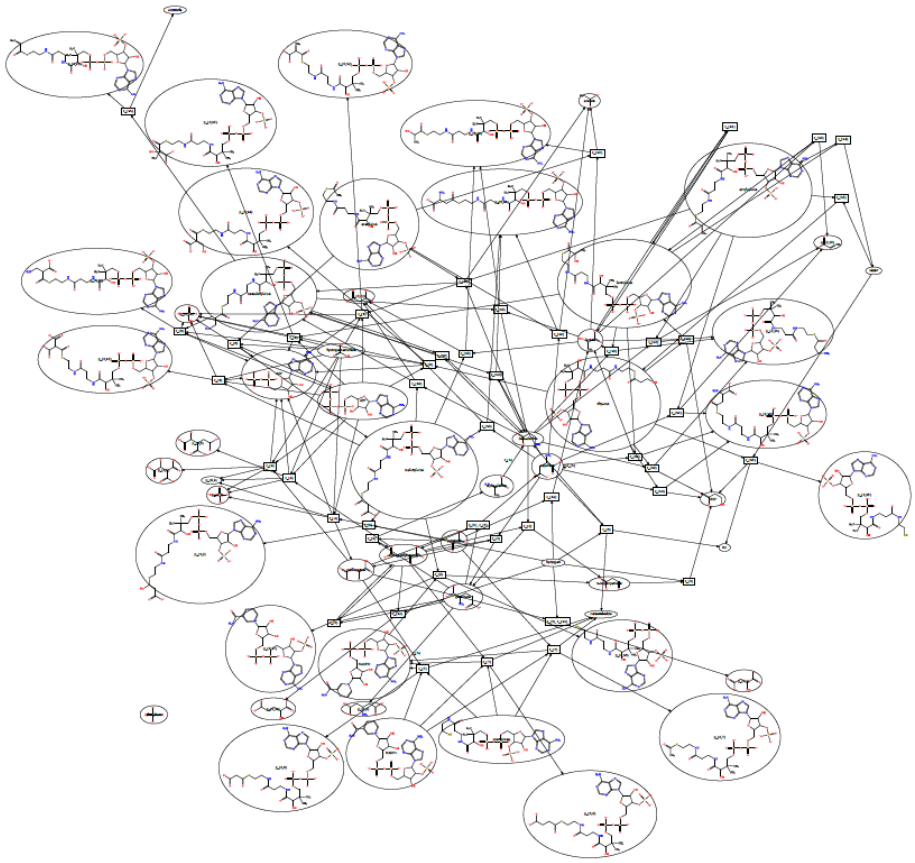


Figure 3.21: The generated directed hypergraph is shown, where the nodes correspond to the chemical species and the directed hyper edges correspond to the reaction, which connect the educts with the products. With the set of maximum rewrite rules and the set of compounds and cofactors all known pathways could be reproduced.

To find novel pathways, which have better or equal advantages than the industrial implemented one, the minimum rewrite rules, which contains just the atoms and bonds of the reaction center, were used as input rules. And also all compounds and cofactors were set as input graphs. And again a network was performed with the network construction tool MØD. However, there was a combinatorial explosion in producing the compounds and reactions, because the reaction rules were too unspecific. In the next run, those minimum rules

were modified. The rules were changed in such a way that a slight expansion of the context was carried out in addition to the reaction centers. In other words, neighbour atoms were partly included in order to make the rule more specific and thus avoid the combinatorial explosion.

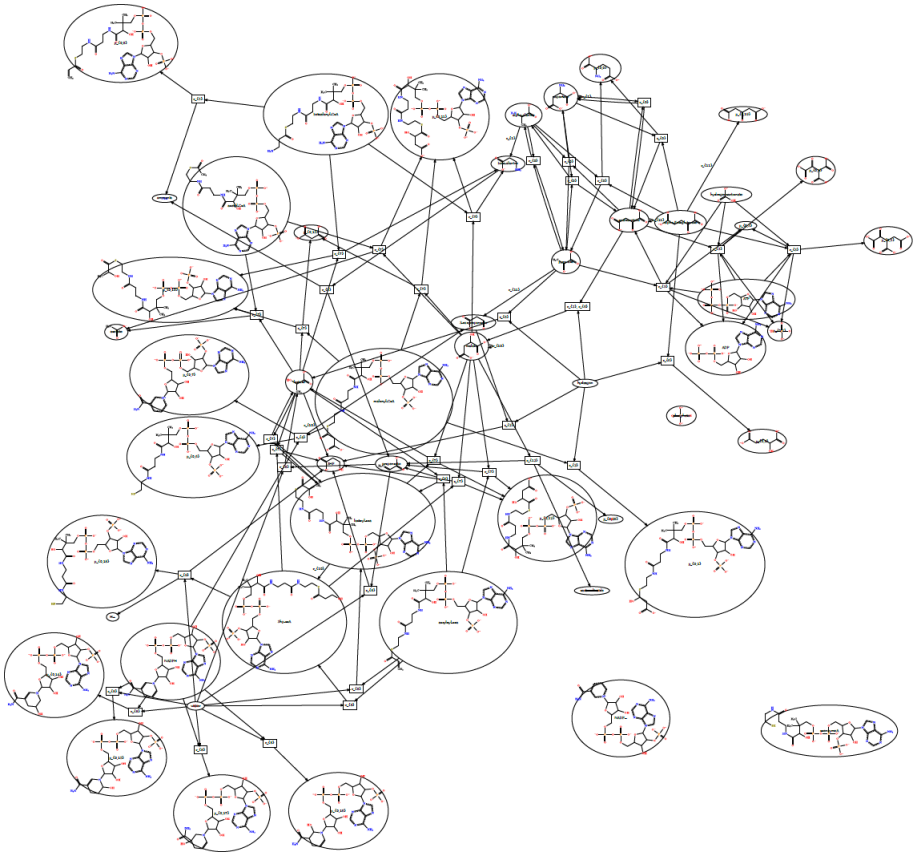


Figure 3.22: The generated directed hypergraph of the novel pathways is illustrated. With the set of minimum rewrite rules and the set of compounds and cofactors novel routes for 3HP from pyruvate could be produced.

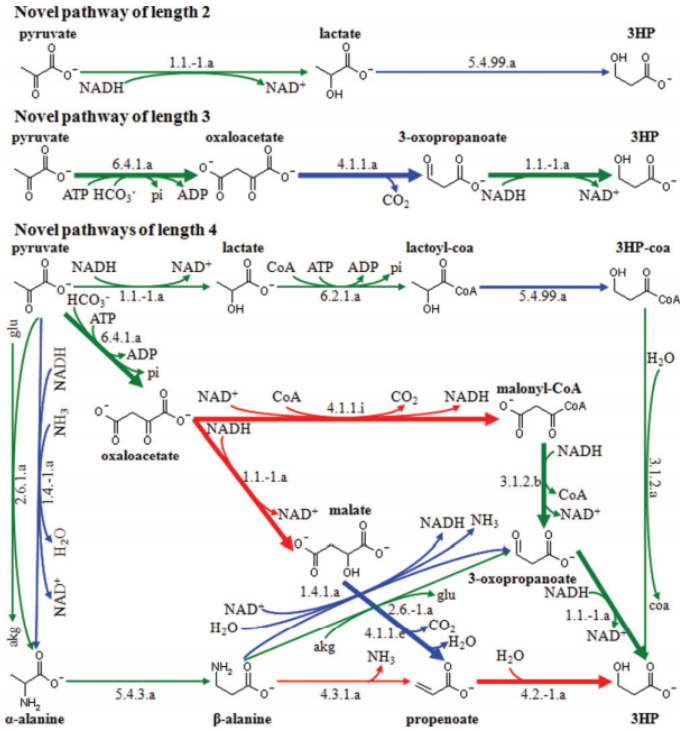


Figure 3.23: The novel reaction pathways lesser than five are listed in this image, which were created by the BNICE framework. [2]

All generated routes contain the same intermediate products as in the known pathways with the exception of malate and propenoate.

Novel generated pathways
Reaction length of 2:
Pyruvate → lactate → 3HP
Pyruvate → 3-oxopropanoate → 3HP

Reaction length of 4:
Pyruvate → lactate → lactoyl-CoA → 3HP-CoA → 3HP
Pyruvate → oxaloacetate → malate → propenoate → 3HP
Pyruvate → α-alanine → β-alanine → propenoate → 3HP

The novel reactions listed above are the same as the paths generated in the BNICE framework with the exception of pyruvate → 3-oxopropanoate → 3HP. All novel reaction listed above were evaluated already in BNICE framework by their thermodynamic feasibility and maximum yield for the cellular production of 3HP. [75] [78] The thermodynamic feasibility of the reaction and routes was assessed based on the mM Gibbs free energy change of reaction. With the thermodynamics based metabolic flux analysis (TMFA) the maximum yield for the production of 3HP with glucose as raw material in E.coli could also be assessed. [75]

[2] [75]

The evaluation showed that the shortest pathways pyruvate → lactate → 3HP has the same maximum yield production as the industrial implemented one. So, there is no ATP consume involved in one of the reaction steps. An advantages over the commercial used pathway is that it has a reaction length of two. So, this pathway makes a promising alternative to the proposed industrial used pathway.
The second shortest pathway pyruvate → 3-oxopropanoate → 3HP has no evaluation, because the BNICE framework did not generate this pathway. So an evaluation has yet to be done to this novel generated pathway.
The reaction pathways with a length of four has as many reaction steps as the implemented one. The reaction paths pyruvate → lactate → lactoyl-CoA → 3HP-CoA → 3HP and pyruvate → oxaloacetate → malate → propenoate → 3HP are less promising, because they have one reaction step, which involves ATP that decrease the maximum yield by half. That makes these two pathways less economical.
The reaction route pyruvate → α-alanine → β-alanine → propenoate → 3HP has almost the same intermediates and biochemical reaction steps as the industrial implemented one. They differ only in the third intermediate, instead of 3-oxopropanoate, propenoate is now produced. But both routes have the same advantages over the thermodynamic feasibility and maximum yield of 3HP.

The industrial implemented one has no clear advantages or disadvantages over the created novel pathways, which makes them a great alternative for the synthesis planning. This approach can be used for other high valuable chemicals, which don't have proposed biosynthesis pathways yet, and this biochemical alternative routes can be used to replace the organic synthesis.

In the case of BNICE framework the pathways were generated and found by using reaction rules, which were based on the third level generalized EC classes, but this has a limitation in their flexibility of changing their specificity of the reaction rules.
In the framework that has been created in this master thesis, there is a possible systematic way that can bring flexibility in terms of finding new biosynthesis routes for high valuable organic compound such as 3HP. At the moment the maximum and minimum rules are saved in the created database ERRD. The maximum rules include all atoms and bonds of a specific reaction, while the minimum rule includes the atoms and bonds of the reaction center. It is theoretically also possible to systematically find all intermediates between the maximum and minimum rules by reducing or expanding the context or the atoms environment of the rewrite rules. In this way, this framework gets much more flexible, when it comes to find new promising synthetic alternatives, which is essential for the further synthesis planning.
If the complete maximum context is utilized for generating networks by using automatic generating tools such as MØD, there will probably no novel pathways been found, because the maximum rules are too specific. If the generated network is fed with the minimal context, of the rewrite rules, which is the most unspecific rule, many possible novel routes can be created, which can lead to a combinatorial explosion. But most of the generated pathways have reaction steps involved, which cannot be chemically catalyzed by the certain enzyme in the organism. These synthesis alternatives are less promising, because mostly would not work in the experimental implementation. The expansion of differ-

ent levels of the context between the maximum and the minimum rewrite rule can lead to an improvement in finding promising biochemical synthetic routes for high valuable organic compounds, which are more likely to be produced by known enzymes in the organism.
Although only the maximum and minimum rewrite rules were created in this thesis, this can be seen as a groundwork for being able to predict chemical possible pathways, which were catalyzed by enzymes, by changing the level of context in a reaction rule.

Another disadvantage of the BNICE framework, which were utilized to produce the pathways, is that the reaction rules don't have atom mappings involved. But the rewrite rules created in this thesis have atom mappings, which can be used to investigate reaction mechanism in a biochemical pathway. With the atom mapping it is clear which transformation can be catalysed by which enzyme. Also, it is important to have atom maps to experimentally corroborate that these predicted pathways can be used when implementing in an organism. Since the atom maps are available, it is possible to construct arbitrary networks including the atom maps.

Bibliography

[1] Julia A Hasler, Ronald Estabrook, Michael Murray, Irina Pikuleva, Michael Waterman, Jorge Capdevila, Vijakumar Holla, Christian Helvig, John R Falck, Geoffrey Farrell, et al. Human cytochromes p450. *Molecular aspects of medicine*, 20(1-2):1–137, 1999.

[2] David C Whitcomb and Mark E Lowe. Human pancreatic digestive enzymes. *Digestive diseases and sciences*, 52(1):1–17, 2007.

[3] Masaaki Kotera, Yasushi Okuno, Masahiro Hattori, Susumu Goto, and Minoru Kanehisa. Computational assignment of the ec numbers for genomic-scale analysis of enzymatic reactions. *Journal of the American Chemical Society*, 126(50):16487–16498, 2004.

[4] Dietmar Schomburg and Ida Schomburg. Enzyme databases. In *Data Mining Techniques for the Life Sciences*, pages 113–128. Springer, 2010.

[5] SC Kou, Binny J Cherayil, Wei Min, Brian P English, and X Sunney Xie. Single-molecule michaelis- menten equations, 2005.

[6] Alexey V Melkikh and Andrei Khrennikov. Molecular recognition of the environment and mechanisms of the origin of species in quantum-like modeling of evolution. *Progress in Biophysics and Molecular Biology*, 130:61–79, 2017.

[7] Daniel E Koshland Jr. Das schüssel-schloß-prinzip und die induced-fit-theorie. *Angewandte Chemie*, 106(23-24):2468–2472, 1994.

[8] D E Koshland. Application of a theory of enzyme specificity to protein synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 44(2):98, 1958.

[9] Lloyd Wolfinbarger Jr. *Enzyme Regulation in Metabolic Pathways*. Wiley Online Library, 2017.

[10] Ida Schomburg, Antje Chang, and Dietmar Schomburg. Brenda, enzyme data and metabolic information. *Nucleic acids research*, 30(1):47–49, 2002.

[11] Peter D Karp, Monica Riley, Suzanne M Paley, and Alida Pellegrini-Toole. The metacyc database. *Nucleic acids research*, 30(1):59–61, 2002.

[12] António J M Ribeiro, Gemma L Holliday, Nicholas Furnham, Jonathan D Tyzack, Katherine Ferris, and Janet M Thornton. Mechanism and catalytic

site atlas (m-csa): a database of enzyme reaction mechanisms and active sites. *Nucleic acids research*, 46(D1):D618–D623, 2018.

[13] Peter D Karp, Monica Riley, Milton Saier, Ian T Paulsen, Suzanne M Paley, and Alida Pellegrini-Toole. The ecocyc and metacyc databases. *Nucleic acids research*, 28(1):56–59, 2000.

[14] Vassily Hatzimanikatis, Chunhui Li, Justin A Ionita, and Linda J Broadbelt. Metabolic networks: enzyme function and metabolite structure. *Current opinion in structural biology*, 14(3):300–306, 2004.

[15] Mario Latendresse, Jeremiah P Malerich, Mike Travers, and Peter D Karp. Accurate atom-mapping computation for biochemical reactions. *Journal of chemical information and modeling*, 52(11):2970–2982, 2012.

[16] Nuno Osório, Paulo Vilaça, and Miguel Rocha. A critical evaluation of automatic atom mapping algorithms and tools. In *International Conference on Practical Applications of Computational Biology & Bioinformatics*, pages 257–264. Springer, 2017.

[17] Ron Caspi, Richard Billington, Carol A Fulcher, Ingrid M Keseler, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Peter E Midford, Quang Ong, Wai Kit Ong, et al. The metacyc database of metabolic pathways and enzymes. *Nucleic acids research*, 46(D1):D633–D639, 2018.

[18] Ida Schomburg, Antje Chang, Sandra Placzek, Carola Söhngen, Michael Rother, Maren Lang, Cornelia Munaretto, Susanne Ulas, Michael Stelzer, Andreas Grote, et al. Brenda in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in brenda. *Nucleic acids research*, 41(D1):D764–D772, 2012.

[19] J Gasteiger. Chemical structure systems. computational techniques for representation, searching, and processing of structural information, edited by je ash, wa warr and p. willett: Ellis horwood, chichester, 1991, 351 pages, price us $67.50, isbn0 - 13 - 126699 - 3, 1993$.

[20] S Grimme. Wiley interdiscip. rev.: Comput. mol. sci. 1, 211 (2011).

[21] Pieter P Plehiers, Guy B Marin, Christian V Stevens, and Kevin M Van Geem. Automated reaction database and reaction network analysis: extraction of reaction templates using cheminformatics. *Journal of cheminformatics*, 10(1):1–18, 2018.

[22] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

[23] Wikipedia. Simplified molecular-input line-entry system.

[24] David Weininger, Arthur Weininger, and Joseph L Weininger. Smiles. 2. algorithm for generation of unique smiles notation. *Journal of chemical information and computer sciences*, 29(2):97–101, 1989.

[25] Inc. Daylight Chemical Information System. Smiles - a simplified chemical language. URL:`https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html`.

[26] Open Smiles Project.

[27] Daylight. Reaction smiles. URL:https://www.daylight.com/meetings/summerschool01/course/basics/smirks.html.

[28] Inc. Daylight Version 4.9 Daylight Chemical Information Systems. Daylight theory manual.

[29] Nina Jeliazkova and Nikolay Kochev. Ambit-smarts: Efficient searching of chemical structures and fragments. *Molecular informatics*, 30(8):707–720, 2011.

[30] Richard GA Bone, Michael A Firth, and Richard A Sykes. Smiles extensions for pattern matching and molecular transformations: Applications in chemoinformatics. *Journal of chemical information and computer sciences*, 39(5):846–860, 1999.

[31] Robert Schmidt, Emanuel SR Ehmki, Farina Ohm, Hans-Christian Ehrlich, Andriy Mashychev, and Matthias Rarey. Comparing molecular patterns using the example of smarts: Theory and algorithms. *Journal of chemical information and modeling*, 59(6):2560–2571, 2019.

[32] Stephen R Heller, Alan McNaught, Igor Pletnev, Stephen Stein, and Dmitrii Tchekhovskoi. Inchi, the iupac international chemical identifier. *Journal of cheminformatics*, 7(1):23, 2015.

[33] John W Raymond and Peter Willett. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *Journal of computer-aided molecular design*, 16(7):521–533, 2002.

[34] Stephen Heller, Alan McNaught, Stephen Stein, Dmitrii Tchekhovskoi, and Igor Pletnev. Inchi-the worldwide chemical structure identifier standard. *Journal of cheminformatics*, 5(1):7, 2013.

[35] the free encyclopedia Wikipedia. International chemical identifier.

[36] inchi.info. Inchikey. URL:http://inchi.info/inchikey_overview_en.html.

[37] Arthur Dalby, James G Nourse, W Douglas Hounshell, Ann KI Gushurst, David L Grier, Burton A Leland, and John Laufer. Description of several chemical structure file formats used by computer programs developed at molecular design limited. *Journal of chemical information and computer sciences*, 32(3):244–255, 1992.

[38] Reiko Heckel. Graph transformation in a nutshell. *Electronic notes in theoretical computer science*, 148(1):187–198, 2006.

[39] V Yegnarayanan and Gulshan Wadhwa. Graph transforms for modeling chemical reaction pathways. *International Journal of Bioinformatics and Biological Science*, 1(1):109–128, 2013.

[40] Jakob L Andersen, Christoph Flamm, Daniel Merkle, and Peter F Stadler. Generic strategies for chemical space exploration. *arXiv preprint arXiv:1302.4006*, 2013.

[41] Jakob L Andersen, Christoph Flamm, Daniel Merkle, and Peter F Stadler. A software package for chemically inspired graph transformation. In *International Conference on Graph Transformation*, pages 73–88. Springer, 2016.

[42] Andrea Corradini, Ugo Montanari, Francesca Rossi, Hartmut Ehrig, Reiko Heckel, and Michael Löwe. Algebraic approaches to graph transformation–part i: Basic concepts and double pushout approach. In *Handbook Of Graph Grammars And Computing By Graph Transformation: Volume 1: Foundations*, pages 163–245. World Scientific, 1997.

[43] Jakob L Andersen, Christoph Flamm, Daniel Merkle, and Peter F Stadler. Rule composition in graph transformation models of chemical reactions. *Match*, 80(3):661–704, 2018.

[44] Christoph Flamm and Martin Mann. Ggl tutorial: Graph rewrite rules [online]. institute for theoretical chemistry, univarsity of vienna and bioinformatics group, university of freiburg. URL:`http://www.tbi.univie.ac.at/software/GGL/Tutorials/tutorial-rules.pdf`, 2013.

[45] Rolf Fagerberg, Christoph Flamm, Daniel Merkle, Philipp Peters, and Peter F Stadler. On the complexity of reconstructing chemical reaction networks. *Mathematics in Computer Science*, 7(3):275–292, 2013.

[46] Jakob L Andersen, Christoph Flamm, Daniel Merkle, and Peter F Stadler. Inferring chemical reaction patterns using rule composition in graph grammars. *Journal of Systems Chemistry*, 4(1):4, 2013.

[47] Steffen Klamt, Utz-Uwe Haus, and Fabian Theis. Hypergraphs and cellular networks. *PLoS computational biology*, 5(5), 2009.

[48] Jonathan Laurent, Jean Yang, and Walter Fontana. Counterfactual resimulation for causal analysis of rule-based models. In *IJCAI*, pages 1882–1890, 2018.

[49] J.L. Andersen, C. Flamm, D. Merkle, and P.F. Stadler. 50 shades of rule composition from chemical reactions to higher levels of abstraction. volume 8738, pages 117–135. Springer Verlag, 2014.

[50] Ari Rantanen, Juho Rousu, Paula Jouhten, Nicola Zamboni, Hannu Maaheimo, and Esko Ukkonen. An analytic and systematic framework for estimating metabolic flux ratios from 13 c tracer experiments. *BMC bioinformatics*, 9(1):266, 2008.

[51] Juho Rousu, Ari Rantanen, Hannu Maaheimo, Esa Pitkänen, Katja Saarela, and Esko Ukkonen. A method for estimating metabolic fluxes from incomplete isotopomer information. In *International Conference on Computational Methods in Systems Biology*, pages 88–103. Springer, 2003.

[52] William Lingran Chen, David Z Chen, and Keith T Taylor. Automatic reaction mapping and reaction center detection. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 3(6):560–593, 2013.

[53] JM Harrison and MF Lynch. Computer analysis of chemical reactions for storage and retrieval. *Journal of the Chemical Society C: Organic*, (15):2082–2087, 1970.

[54] Michael F Lynch and Peter Willett. The production of machine-readable descriptions of chemical reactions using wiswesser line notations. *Journal of Chemical Information and Computer Sciences*, 18(3):149–154, 1978.

[55] Hans-Christian Ehrlich and Matthias Rarey. Maximum common subgraph isomorphism algorithms and their applications in molecular science: a review.

Wiley Interdisciplinary Reviews: Computational Molecular Science, 1(1):68–79, 2011.

[56] Lingran Chen. Substructure and maximal common substructure searching. In *Computational Medicinal Chemistry for Drug Discovery*, pages 509–540. CRC Press, 2003.

[57] Harry L Morgan. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of Chemical Documentation*, 5(2):107–113, 1965.

[58] Michael F Lynch and Peter Willett. The automatic detection of chemical reaction sites. *Journal of Chemical Information and Computer Sciences*, 18(3):154–159, 1978.

[59] James J McGregor. Backtrack search algorithms and the maximal common subgraph problem. *Software: Practice and Experience*, 12(1):23–34, 1982.

[60] James J McGregor and Peter Willett. Use of a maximum common subgraph algorithm in the automatic identification of ostensible bond changes occurring in chemical reactions. *Journal of Chemical Information and Computer Sciences*, 21(3):137–140, 1981.

[61] Robert Körner and Joannis Apostolakis. Automatic determination of reaction mappings and reaction center information. 1. the imaginary transition state energy approach. *Journal of chemical information and modeling*, 48(6):1181–1189, 2008.

[62] Clemens Jochum, Johann Gasteiger, and Ivar Ugi. The principle of minimum chemical distance (pmcd). *Angewandte Chemie International Edition in English*, 19(7):495–505, 1980.

[63] Rina Dechter and Judea Pearl. Generalized best-first search strategies and the optimality of a. *Journal of the ACM (JACM)*, 32(3):505–536, 1985.

[64] Russell Stuart, Norvig Peter, et al. Artificial intelligence: a modern approach, 2003.

[65] Markus Heinonen, Sampsa Lappalainen, Taneli Mielikäinen, and Juho Rousu. Computing atom mappings for biochemical reactions without subgraph isomorphism. *Journal of Computational Biology*, 18(1):43–58, 2011.

[66] Alexander Schrijver. *Theory of linear and integer programming*. John Wiley & Sons, 1998.

[67] Eric L First, Chrysanthos E Gounaris, and Christodoulos A Floudas. Stereochemically consistent reaction mapping and identification of multiple reaction mechanisms through integer linear optimization. *Journal of chemical information and modeling*, 52(1):84–92, 2012.

[68] German A Preciat Gonzalez, Lemmer RP El Assal, Alberto Noronha, Ines Thiele, Hulda S Haraldsdóttir, and Ronan MT Fleming. Comparative evaluation of atom mapping algorithms for balanced metabolic reactions: application to recon 3d. *Journal of cheminformatics*, 9(1):39, 2017.

[69] Syed Asad Rahman, Gilliean Torrance, Lorenzo Baldacci, Sergio Martínez Cuesta, Franz Fenninger, Nimish Gopal, Saket Choudhary, John W

May, Gemma L Holliday, Christoph Steinbeck, et al. Reaction decoder tool (rdt): extracting features from chemical reactions. *Bioinformatics*, 32(13):2065–2066, 2016.

[70] Syed Asad Rahman, Sergio Martinez Cuesta, Nicholas Furnham, Gemma L Holliday, and Janet M Thornton. Ec-blast: a tool to automatically search and compare enzyme reactions. *Nature methods*, 11(2):171–174, 2014.

[71] Hans Kraut, Josef Eiblmaier, Guenter Grethe, Peter Low, Heinz Matuszczyk, and Heinz Saller. Algorithm for reaction classification. *Journal of chemical information and modeling*, 53(11):2884–2895, 2013.

[72] Akhil Kumar and Costas D Maranas. Clca: maximum common molecular substructure queries within the metrxn database. *Journal of chemical information and modeling*, 54(12):3417–3438, 2014.

[73] Thomas G Kristensen, Jesper Nielsen, and Christian NS Pedersen. A tree-based method for the rapid screening of chemical fingerprints. *Algorithms for Molecular Biology*, 5(1):9, 2010.

[74] Darko Butina. Unsupervised data base clustering based on daylight's fingerprint and tanimoto similarity: A fast and automated way to cluster small and large data sets. *Journal of Chemical Information and Computer Sciences*, 39(4):747–750, 1999.

[75] Christopher S Henry, Linda J Broadbelt, and Vassily Hatzimanikatis. Discovery and analysis of novel metabolic pathways for the biosynthesis of industrial chemicals: 3-hydroxypropanoate. *Biotechnology and bioengineering*, 106(3):462–473, 2010.

[76] Karin Hofvendahl and Bärbel Hahn-Hägerdal. Factors affecting the fermentative lactic acid production from renewable resources1. *Enzyme and microbial technology*, 26(2-4):87–107, 2000.

[77] Alvin L Young et al. Biotechnology for food, energy, and industrial products: new opportunities for bio-based products. *Environmental Science and Pollution Research*, 10(5):273–276, 2003.

[78] Elizabeth Brunk, Marilisa Neri, Ivano Tavernelli, Vassily Hatzimanikatis, and Ursula Rothlisberger. Integrating computational methods to retrofit enzymes to synthetic pathways. *Biotechnology and Bioengineering*, 109(2):572–582, 2012.

[79] Patrick F Suthers and Douglas C Cameron. Production of 3-hydroxypropionic acid in recombinant organisms, February 8 2005. US Patent 6,852,517.

[80] Ravi R Gokarn, Olga V Selifonova, Holly Jean Jessen, Steven John Gort, Thorsten Selmer, and Wolfgang Buckel. 3-hydroxypropionic acid and other organic compounds, March 6 2007. US Patent 7,186,541.

I have tried to locate all owners of the image rights and obtained their consent to the use of the images in this work. Should a copyright infringement become known, I request that you notify me.