UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Expanded natural product chemspace

In silico expansion of LOTUS with Pickaxe

Student: Pascal Amrein
SIUS number: 16-164-006
Group: COMMONS Lab
Supervisor: Pierre-Marie Allard

Hand-in: 3. September 2024

Contents

Abstract **1**

1 Introduction **2**

2 Materials and Methods **3**

2.1 LOTUS dataset 3

2.2 Dataset-extractor 6

2.3 MINEs & Pickaxe 6

2.4 Exploration of LOTUS 13

2.5 Dataprocessing 15

3 Results **16**

3.1 Ruleset 16

3.2 MINES 17

4 Discussion **24**

5 References **25**

5.1 Abbreviations 25

5.2 Bibliography 25

5.3 Figures 27

5.4 Tables 27

5.5 Links 29

Appendix **29**

Declaration of Authenticity 29

Abstract

Exploring the chemical space of natural products (NPs) is crucial for the discovery of new bioactive compounds. This study expands the LOTUS dataset, a comprehensive database of over 750,000 structural organism pairs, by using Pickaxe software to predict novel metabolic reactions and compounds. After selecting unique starting compounds and removing stereochemistry, approximately 140,000 molecules from LOTUS were available for prediction with Pickaxe, resulting in the generation of 3.2 million new potential molecules and 3.4 million new reactions based on 250 chemical rules. For the input of molecules, SMILES (Simplified Molecular Input Line Entry System) was used to represent chemical structures and SMARTS (SMiles ARbitrary Target Specification) was used to encode the chemical transformation rules. The methodology focused on the use of in silico techniques to generate these predictions, which were stored and analysed in a Mongo database. A notable result was the frequent prediction of bromine, with bromine being the most common element among the newly predicted molecules. This frequent occurrence can be attributed to the versatility of bromine in the formation of stable compounds and its known role in biological systems, e.g. as a cofactor in enzymatic reactions. Principal component analysis (PCA) showed a broad distribution of the predicted compounds around the starting compounds, indicating the plausibility of the reactions used. These results emphasise the potential of computational tools in expanding the chemical space of natural products and represent a valuable resource for future drug discovery. As a next step, the generation of mass spectrometry (MS) spectra for the predicted compounds and their comparison with experimentally generated spectra by untargeted metabolomics could help in the discovery of new molecules.

Keyword: *in silico, expanded chemical structure, natural products*

1 Introduction

The simplest definition for a natural product (NP) is a small molecule that is produced by a biological source. ("All Natural" 2007) The resulting discussion about primary metabolites (vital) and secondary metabolites (protective functions and co) is not always clear.

Natural products research focuses on the study and utilization of chemical compounds produced by living organisms, including plants, microbes, and marine life. These compounds often have unique and complex structures, making them valuable for drug discovery and other applications in medicine and chemistry. (Chainani et al. 2023)

Without the ability to access and process these disparate NP data points, information becomes fragmented, hindering scientific progress (Balietti, Mäs, and Helbing 2015). In this context, contemporary bioinformatic tools facilitate the (re-)interpretation and (re-)annotation of (existing) datasets that document the molecular aspects of biodiversity (Mongia and Mohimani 2021).

The LOTUS initiative describes this connection with over 750,000 referenced structure-organsimus pairs. An additional and fundamental element of a structure-organism pair is a reference to the experimental evidence that establishes the linkages between the chemical structure and the biological organism. (Rutz et al. 2022)

An important aspect of the many discoveries of new natural products in recent years has been made possible by technological progress. Mass spectrometry and next generation sequencing should be mentioned here. (Balietti, Mäs, and Helbing 2015)

Next-Generation Sequencing (NGS) allows researchers to sequence the genomes of organisms that produce natural products. By analysing these genomes, scientists can identify genes and gene clusters responsible for the biosynthesis of natural compounds. This genomic information enables the discovery of new natural products, particularly from uncultivable or rare organisms.

Mass spectrometry (MS) plays a crucial role in natural products research by enabling the precise analysis and identification of complex molecules. MS helps in determining the molecular weight and structure of natural compounds, which is essential for understanding their biological activity. MS is a powerful analytical technique used to identify and quantify molecules based on their mass-to-charge ratio. It allows researchers to analyse minute quantities of compounds with high sensitivity, making it an indispensable tool for discovering new natural products and understanding their chemical properties. MS is also used in the dereplication process to avoid rediscovery of known compounds. (Jarmusch et al. 2020) (Ebbels et al. 2023)

Four points must be emphasized for a good analysis: (1) experimental design, (2) pre-analytical (sample collection and preparation), (3) analytical (chromatography and detection), and (4) post-analytical (data processing). In this work, the 4th point (data processing) is mainly emphasized.

The “targeted mass spectrometry” method is very frequently used in analyses. Here, the spectra generated by the MS are searched for specific “peaks” that are known. This requires a database to find and quantify the peaks.

A more sophisticated procedure is “untargeted mass spectrometry,” in which the specific targets are not predetermined. The aim here is to discover and identify as many molecules as possible without prior knowledge or bias towards certain compounds. This method is used more frequently in proteomics and metabolomics. This method is more time and resource consuming as it requires extensive data processing and validation.

In order to find possible new molecules, it can be helpful to already have an idea of their structure. This is where the in silico (computational) approach can help. New molecules are predicted on the basis of known chemical rules and known molecules. The resulting reaction networks can be used for multiple applications such as designing novel biosynthetic pathways and annotating untargeted metabolomics data.

In this master's thesis, the LOTUS database was expanded with approx. +140,000 molecules, so that in the end 3.2 million new possible molecules were predicted using approx. 250 chemical rules.

Pickaxe was used for the expansion of the lotus molecules.

Pickaxe predicts novel metabolic reactions and compounds that can be used for a variety of applications. This software is open source and available as part of the Python package MINE Database (https://pypi.org/project/minedatabase/) or on GitHub (https://github.com/tyo-nu/MINE-Database). With the help of High Performance Computers, this task was carried out and discussed further here.

2 Materials and Methods

The molecule expansion was carried out on the basis of the LOTUS dataset with 220’834 unique compounds. This was expanded with 250 chemical rules and a specific list of 33 coreactants. To accomplish this task, we first practiced with a small dataset generated with the dataset extractor to validate the pipeline in less time. Pickaxe removes the stereochemistry of the input file and ended up with 147’861 compounds. Pickaxe expanded the chemical space using this input file and the rulset. These components were then stored in a MongoDB.

This database was then used for analysis.

2.1 LOTUS dataset

Paper (DOI:70780): https://elifesciences.org/articles/70780
latest LOTUS dataset (v10): https://zenodo.org/records/7534071

LOTUS initiative: https://lotus.nprod.net/
website (old): https://lotus.naturalproducts.net

The LOTUS dataset contains information about organisms and their discovered chemical molecules.
The latest LOTUS dataset (v10) has 792'364 entries with 39 columns divided into chemical structure, natural products (NP), organism and sources.

With the example of the molecule "Limonene" this division of the chemical structure (Table 1), the organism structure (Table 2) and source (Table 3) is shown.
Limonen has 1'176 entries in LOTUS. In the following example we concentrate only on one entry.

col nr.	**Column name**	**Example "Limonene"**
1	structure_wikidata	http://www.wikidata.org/entity/Q278809
2	structure_inchikey	XMGQYMWWDOXHJM-UHFFFAOYSA-N
3	structure_inchi	InChI=1S/C10H16/c1-8(2)10-6-4-9(3)5-7-10/h4,10H,1,5-7H2,2-3H3
4	structure_smiles	C=C(C)C1CC=C(C)CC1
5	structure_molecular_formula	C10H16
6	structure_exact_mass	136.125200512
7	structure_xlogp	3.3089
8	structure_smiles_2D	C=C(C)C1CC=C(C)CC1
9	structure_cid	22311
10	structure_nameIupac	1-methyl-4-prop-1-en-2-ylcyclohexene
11	structure_nameTraditional	Limonene
12	structure_stereocenters_total	1
13	structure_stereocenters_unspecified	1
14	structure_taxonomy_npclassifier_01pathway	Terpenoids
15	structure_taxonomy_npclassifier_02superclass	Monoterpenoids
16	structure_taxonomy_npclassifier_03class	Menthane monoterpenoids
17	structure_taxonomy_classyfire_chemontid	1401
18	structure_taxonomy_classyfire_01kingdom	Organic compounds
19	structure_taxonomy_classyfire_02superclass	Lipids and lipid-like molecules
20	structure_taxonomy_classyfire_03class	Prenol lipids
21	structure_taxonomy_classyfire_04directparent	Menthane monoterpenoids

Table 1: LOTUS dataframe columns - chemical structure

col nr.	Column name	Example "Limonene"
22	organism_wikidata	http://www.wikidata.org/entity/Q26726
23	organism_name	Cannabis sativa
24	organism_taxonomy_gbifid	5361880
25	organism_taxonomy_ncbiid	3483
26	organism_taxonomy_ottid	84004
27	organism_taxonomy_01domain	Eukaryota
28	organism_taxonomy_02kingdom	Archaeplastida
29	organism_taxonomy_03phylum	Streptophyta
30	organism_taxonomy_04class	Magnoliopsida
31	organism_taxonomy_05order	Rosales
32	organism_taxonomy_06family	Cannabaceae
33	organism_taxonomy_07tribe	
34	organism_taxonomy_08genus	Cannabis
35	organism_taxonomy_09species	Cannabis sativa
36	organism_taxonomy_10varietas	

Table 2: LOTUS dataframe columns - organism structure

col nr.	Column name	Example "Limonene"
37	reference_wikidata	http://www.wikidata.org/entity/Q40284493
38	reference_doi	10.1021/NP50008A001
39	manual_validation	

Table 3: LOTUS dataframe columns - reference

Under the category natural products ("structure_taxonomy_npclassifier" - column 14-16 in the LOTUS df), terpenoids are the most common (= 296'899). Alkaloids are in third place with 113,284. The chemical structure ("structure_taxonomy_classyfire" - column 18-21 in the LOTUS df) contains mainly organic compounds (organic compounds = 788'301). Of these, "lipids and lipid-like molecules" are very frequently represented.

Among the organisms ("structure_taxonomy_classyfire" - column 27-36 in the LOTUS df), eukaryotes are the most frequent domain with 735'225. Nevertheless, the most frequently mentioned organism name (column 23 in the LOTUS df) is "Streptomyces" (bacteria) with 6547 entries, followed by Homo sapiens with 3'310 entries and Arabidopsis thaliana (plant) with 3'085 entries. More about these statistics can be found on GitHub "https://github.com/commons-research/expanded-naturalproduct-chemspace.git".

The Lotus dataset was used as the basis for predicting new molecules. (Rutz et al. 2022)

2.2 Dataset-extractor

Git-Hub: https://github.com/commons-research/dataset-extractor-lotus
Documentation: https://commons-research.github.io/dataset-extractor-lotus/

The dataset extractor can be used to download the desired LOTUS dataset and sample it from any LOTUS dataset.
For sampling, we have first do choose the taxon level (column 27 - 36 of LOTUS dataset). In this taxon a specific representative can be choosen. For example on the taxonlevel "domain" can be choosen "Eukaryota". From there the sample number can be provided and it will sample it from the given information. This list can either be saved in MINEs format (standard "structure_inchikey" and "structure_smiles") or in the original format (whole columns). More information provided on Git-Hub.

The advantage of handling smaller amounts of data is the reduction in process duration. This allows a "pipeline" (processing of the data) to be set up and checked for errors before larger amounts of data are processed.
Nevertheless, care should be taken to ensure that processing is carried out in parallel for larger datasets. Either this is taken into account during programming (multiprocessing - fully automatic) or the file is split and processed in portions (semi-automatic).

Both were used in this master's thesis. Splitting the file was needed at the beginning because the script (pickaxe_pamrein.py) showed no improvement even when working with multiprocessor. By using the terminal command, the desired number of processes can be specified directly with the keyword "–processes", resulting in a shorter runtime.

2.3 MINEs & Pickaxe

Following resources are provided from MINEs (*Metabolic In silico Network Expansions*):
paper MINEs (2015, James G. Jeffryes): https://doi.org/10.1186/s13321-015-0087-1
paper Pickaxe (2023, Kevin M. Shebek): https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-023-05149-8
documentation: https://mine-database.readthedocs.io/en/latest/pickaxe_run.html#coreactant-and-rule-lists
github: https://github.com/tyo-nu/MINE-Database
website: https://minedatabase.mcs.anl.gov/#/home

MINEs was created with 198 generalised chemical reaction rules constructed with help of BNICE (Biochemical Network Integrated Computational Explorer) in a given source database. It can be choosen from 3 different databases: KEGG, EcoCyc and YMDB. For example from the Kegg (Kyoto Encyclopedia of Genes and Genomes) given 13'307 compounds and predicted 571'368 compounds (fold increase of 43), where 6.99% of the new compounds were found in PubChem. This in silico (only computation) method generated predicted compounds and predicted reactions, which can be used for untargeted metabolomics (unknown compounds). MINEs can be accessed through the website (https://minedatabase.mcs.anl.gov/) or there API. It can be used to find metabolomics search (MS Adduct or MS/MS) or structure search. (Jeffryes et al. 2015)

With Pickaxe can be predict compounds and reactions. For this it is necessary to enter the desired molecule structure (SMILES) and a corresponding ID. The applied reaction rules (SMARTS) can be added by the user or the ones provided by Pickaxe (e.g. JN1224min ruleset derived from MetaCyc) can be used. The ruleset JN1224min (generalised_metacyc) consists of 1224 SMARTS which cover the largest part of the KEGG and BRENDA reactions with the fewest reaction rules. In the first step, the given components are entered and expanded with the reaction rules (Network expansion with RDKit). The filters can be used to "reduce" the predicted molecules.
After these two steps, they can be repeated with the newly generated molecules to generate the next generation or the generated molecules can be saved. For storing the data on the harddisc, it can either use MongoDB or a *.csv file for saving. (Shebek et al. 2023)

2.3.1 Input files

The input file must be a csv-file (comma separated) with the column names "ID" and "SMILES". The SMILES were taken from the column "structure_smiles" of the LOTUS dataset. The ID can be freely selected, but should not occur twice. In this case, "structure_wikidata" was used initially and later "structure_inchikey" as the ID. The LOTUS input file has 220'823 unique entries (structure_smiles).
For the rule set (reaction rules) can be used the existing ones (metacyc generalized, metacyc intermediate, enyzmatic reactions...) In addition to the rule set, other chemical coreactants (for example water) were also provided, which were used for the reactions.

When reading in the input file (Inchykey and SMILE), 220'733 compounds were loaded, with 147'885 remaining (66.9%) after removing the stereochemistry. For the final run the file "EnzymaticCoreactants.tsv" was used with 33 compounds and "EnzymaticReactionRules.tsv" with 250 rules.

The output during the expansion with MINES is shown in the figure 1.

```
----------------------------------------
Initializing pickaxe object

Done intializing pickaxe object
----------------------------------------
Warning: could not load compound: O=c1nc(=O)[nH]c(NC2CC2)[nH]1
Warning: could not load compound: Cc1[nH]c2nc(=N)nc(O)c2[nH]c1=O
...
220733 compounds loaded...
(147885 after removing stereochemistry)
----------------------------------------
Expanding Generation 1

Generation 1: 0 percent complete
...
Generation 1: 100 percent complete
Generation 1 finished in 3932.657823562622 s and contains:
                3284394 new compounds
                3435078 new reactions

Done expanding Generation: 1.
----------------------------------------


----------------------------------------
Writing results to lotus_mines_enzymatic Database
----------------------------------------

--------------- Reactions --------------
Writing Reactions: Chunk 0 of 344
...
Writing Reactions: Chunk 340 of 344
Wrote Reactions in 58.84046483039856 seconds.
----------------------------------------

--------------- Compounds --------------
Writing Compounds: Chunk 0 of 344
...
Writing Compounds: Chunk 340 of 344
Wrote Compounds in 373.1238865852356 seconds.
----------------------------------------

No targets to write to MINE.
-------------- Operators ---------------
Done with Operators Overall--took 1.499133586883545 seconds.
----------------------------------------

--------------- Indices ----------------
Built Indices--took 35.42222213745117 seconds.
----------------------------------------

--------------- Overall ----------------
Execution took 6758.306858301163 seconds.4126091003 sec
```

Figure 1: comprehensed output of mines

2.3.2 Removing the stereochemistry

When the 220'733 compounds are loaded into pickaxe, only 147'861 compounds (66.9 %) will be found in the MINES MongoDB. This compounds are marked with the type = "Starting Compounds" in the MongoDB.

This reduction happens because of the removal of the stereochemistry from the Input SMILES. Some examples will be shown here as a comparison.
Instead of using the "structure_smiles" (3D SMILES) from the LOTUS dataset, we could use the "structure_smiles_2D" with 148'190 unique entries. We did not do that, because we tried to work with the more accurate data.

2.3.2.1 Comparing InChIKey From the 147'861 compounds are 7'858 compounds (5.32 %) found, where the InChIKey changed from the original InChIKey (Inputfile ID) to the stored InChIKey in the MongoDB. Pickaxe "cleanes up" somehow the compounds and also changes their description as an InChIKey. 7'593 compounds (96.6 %) from this 7'858 compounds have a change in the second block of the InChIKey to the sequence "UHFFFAOYSA". This block is typical for flattend stereochemistry of the compound.

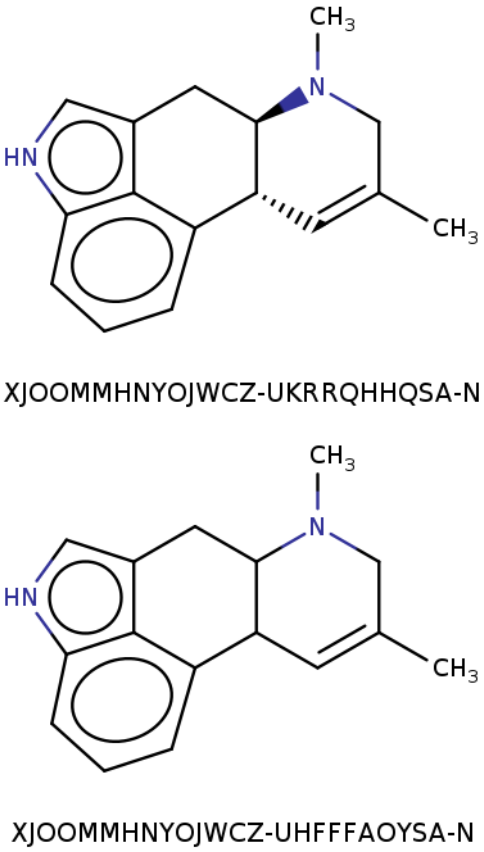

Figure 2: Input SMILES (3D - above) get flattened (down) through the "cleaning" process of pickaxe

The other 265 InChIKey change the last character from "N" (Standard InChIKey) to "O" (Non-standard InChIKey) (example of namechange in inchy is at the tabel 4).

Source	**InChIKey**
MINES Input	LMYYUVAVNWPEPG-GXAHKNHCSA-N
MongoDB	LMYYUVAVNWPEPG-GXAHKNHCSA-O

Table 4: change in the InChIKey N->O

An example can be found in figure 3. This example shows, that the SMILES gives more information than the InchIKey.
The SMILES from the Inputfile is
"...CCC(=O)OC [C@H](COP(=O)(O)OCC [N+](C)(C)C)OC(=O)CCC..."
but after the "cleaning" of pickaxe it is
"...CCC(=O)OC [C@H](COP(=O)([O-])OCC[N+](C)(C)C)OC(=O)CCC...".

As a molecule structure it would look like as the following figure 3.

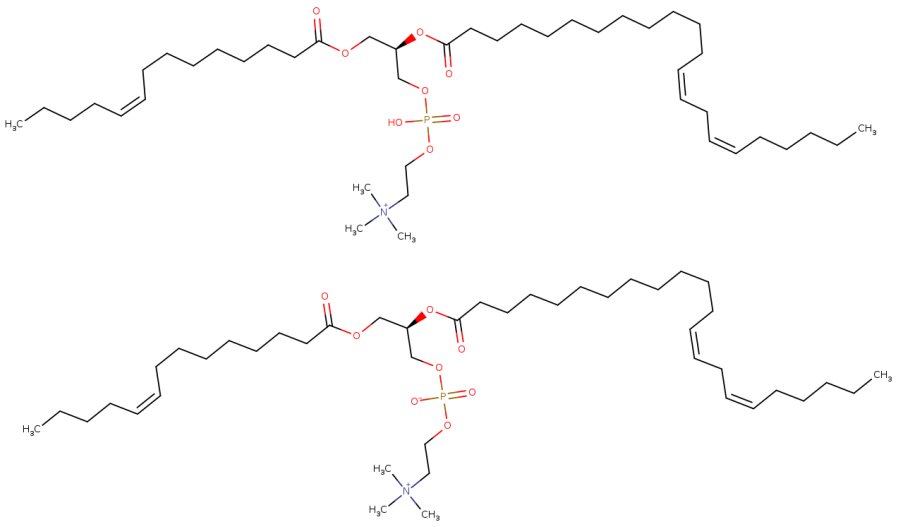

Figure 3: from the inputfile (above) to the stored "starting compound" (down) / change in InChIKey N->O

2.3.3 Ruleset

The rules for the chemical reactions were written as SMARTS accortingly to the daylight specifications. Pickaxe provides several rulesets, two of which were looked at more closely. The generalised one consists of 1224 rules (JN1224min) and the intermediate one of 7358 rules. For this master's thesis, mainly the generalised ruleset was used, which covers a similar range of predicted molecules with fewer rules.
In addition, the "./mine_database/data/original_rules/EnzymaticReactionRules.tsv" was used with 250 reactions and the corresponding "./mine_database/data/original_rules/EnzymaticCoreactants.tsv" file with 33 coreactants. These files were used in earlier versions of Mines and were provided by Pierre-Marie Allard at that time. More about this are in results.

EnzymaticReactionRules.tsv - 250 lines (unique SMARTS)
metacyc_intermediate_rules.tsv - 7359 lines (unknown source)
metacyc_generalized_rules.tsv - 1225 lines (JN1224min from https://doi.org/10.1016/j.ymben.2021.02.006)

unique “SMARTS” in **generalised** rules: 1195

unique “SMARTS” in **intermediate** rules: 7355

same “SMARTS” in **generalised** & **intermediate** rules: 31

A rulefile (*.tsv) must have as a column: Name, Reactants, SMARTS, Products, Other lines such as Comments, counts, Uniprot etc. are not taken into account by Pickaxe itself and are only of interest as additional information. In the **generalized** ruleset, the Uniprot number for the reaction can be found in the “comments” column.
In the **intermediate** ruleset are more information stored then in the generalized. Although these also refer to other databases such as Metacyc, Brenda, kegg. In the **EnzymaticReactionsRules** the information is on other components in the “Comments” column.

The most important part of the rulefile is the line “SMARTS”. It describes the possible reactions for possible molecules.

Example: **SMART in the rule file (rule0402 - metacyc generalized):**
[#6:1]-[#6:2].[#8:3].[#8:4]-[#15:5]»[#6:1].[#6:2]-[#8:4].[#8:3]-[#15:5]

Using the example of limonene, a SMILES reaction looks like this:
C=C(C)C1CC=C(C)CC1.O=P(O)(O)O.Nc1ncnc2c1ncn2C1OC(COP(=O)(O)OP(=O)(O)O)C(O)C1O
»
C.C=C(O)C1CC=C(C)CC1.Nc1ncnc2c1ncn2C1OC(COP(=O)(O)OP(=O)(O)OP(=O)(O)O)C(O)C1O

The SMART in the ruleset shows all possible chemical compounds that must be present for the conversion to work. It does not matter where exactly the breakage and new fusion takes place. The important thing is that the atomic breaks happen according to the SMARTS. The image 4 shows graphically how this can look using the example of limonene as an input molecule.

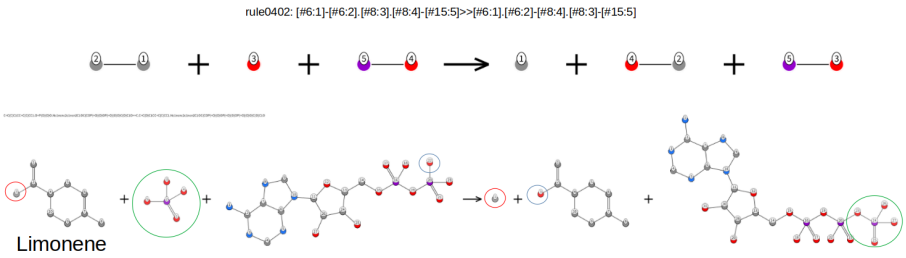

Figure 4: description for the reaction rule0402

It is difficult to understand exactly how the rulesets were compiled, as this was not described in detail in the paper. It was mentioned that BNICE (computational tool to explore the theoretical space of biochemistry - https://lcsb-databases.epfl.ch/) was helpful for this. Registration is required for this.

2.3.3.1 Rule set with Marvin Scetch For creating their own rules, Marvin Scetch (https://docs.chemaxon.com/display/lts-europium/introduction-to-marvinsketch.md) can assist. Marvin Scetch is offered by chemaxon, which also offers other software for chemistry. The program is very intuitive and easy to understand. For example, with simply copy a SMILE past it into the empty window will show the chemical structure.

For practice purposes, we have tried to carry out our own reaction. As an example, we use the decarboxylation reaction as an example (see figure 5). This reaction can be drawn with Marvin Scetch and also checked to see if the reaction is in an equilibrium (check structure with Ctrl + R).

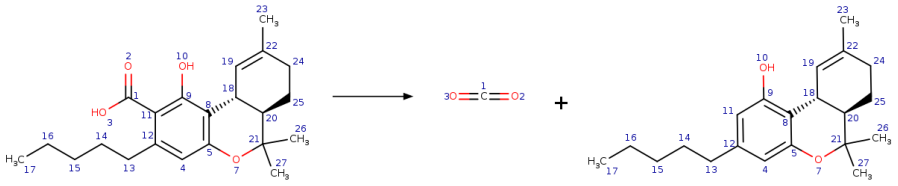

Figure 5: Example of decarboxylation with THC-9

The SMART in Figure 5 will not work with Pickaxe. The main reason for this is that Pickaxe will try to adapt the given structure to the entered compounds. Instead, we should apply the simple rule as shown in Figure 6. It is important to also map the atoms with Marvin Scetch. Otherwise it will not work.

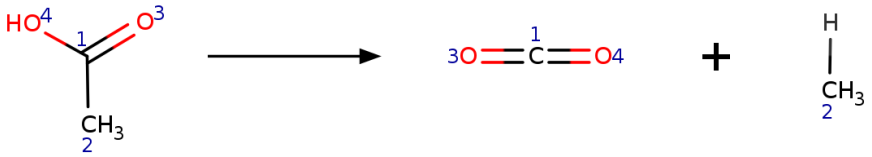

Figure 6: decarboxylation reaction for pickaxe

The reaction from the figure 5 was saved as a SMART and implemented in the reaction tsv with only this rule. This resulted in 21,663 predicted compounds and 23,406 reactions. The full LOTUS dataset was given as input (147,861 output compounds).

2.3.4 Filters

The filters are applied after each round/generation of the prediction of new molecules. This reduces the "chemical space" and keeps the amount of data produced smaller. No filters were applied for this project, as most of the target compounds require specific filtering. All molecules are important in this master's thesis.

The Pickaxe filters are divided into three categories:

1. compound property filters

- *Molecular Weight Filter*: Removes compounds not in range of a specified MoleWeight (min_MW, max_MW)
- *Atomic Composition Filter*: Only elements specified will be filtered (example: atomic_composition_constraints = {"C": [4, 7], "O": [5, 5]})
- *Thermodynamics Filter*: specification of a pH, Ionic strength, pMg and using these to calculate ΔGr. Reactions are then filtered out based on max ΔGr.

2. filters, which compare compounds to a set of provided targets

- *maximum common substructure (MCS) filter*: applies a strict cutoff to the MCS score of compounds to determine which compounds to expand.
- *Similarity Filtering*: A filter that uses a similarity score (Tanimoto or Dice) to determine compounds to expand.
- *Similarity Sampling*: SimilaritySamplingFilter takes a distribution of similarity scores and uses inverse CDF sampling to select N compounds for further expansion.
- *metabolomics filter*: This filter compares the masses (and optionally, predicted retention times) of MINE compounds against peak masses (and retention times) in a metabolomics dataset.

3. filters, which look at properties of reactions

- *Feasibility Filter*: Checks Feasibility of reaction based on deeplearning approach on enzymatic reactions. (Kim et al. 2021).

2.4 Exploration of LOTUS

2.4.1 Part 1 : Wikidata as an ID

At the beginning of the master's thesis (first part), we mainly worked with the Wikidatalink. This chapter serves as an example of how not to do it. In addition, I worked on a high performance computer (HPC) in bern (IBU) where I did not have admin rights as a user and therefore could not work with MongoDB. The good point of the the IBU cluster is, hat they are a lot of cpu, storage and RAM supplies available.

The template script for Pickaxe was modified and used for the prediction of new molecules. The input of the Wikidata and corresponding SMILES consists of 220'834 entries in total. There are 220'783 unique entries from the column "id" and 220'820 entries from the column "smiles". Because of this inconsistency, the wikidatalinks of different SMILES with the same wikidatalink are numbered to make it easier to trace them later. In the case of different wikidatalinks pointing to the same SMILES, this has been left as it is. This molecule is duplicated (with a different ID). However, as there are only a few, this is less significant.

Example of input file with the same SMILE:

```
http://www.wikidata.org/entity/Q106345659,CN(C)CC(=O)O
http://www.wikidata.org/entity/Q4369,CN(C)CC(=O)O
```

Example of input file with the same wikidatalink, which has been customized:

```
http://www.wikidata.org/entity/Q988591_id1,CC/C=C/[C@@H]1C=CCC=CC1
http://www.wikidata.org/entity/Q988591_id2,CC/C=C\[C@@H]1C=CCC=CC1
```

The input file was split into 50 smaller files with approx. 4500 lines each. A reaction file and a compound file were generated for each input file in MINEs. The output files were saved in *.tsv format. The files were transformed into parquet files in order to save space without any loss of performance. Apache Parquet (https://parquet.apache.org/) is an open source, column-oriented data file format designed for efficient data storage and retrieval. It provides high performance compression and encoding schemes to handle complex data in bulk and is supported in many programming language andanalytics tools.

The compound files could be reduced by approx. 75% storage capacity (4.4Gb to 1Gb) and the reaction files by approx. 78% storage capacity (11.4Gb to 2.5Gb).

2.4.2 Part 2 : InchyKey as an ID

At the end of the master thesis (Part 2), we mainly worked with the Inchy-Keys as an "id" for the input file. The main reason is that the Inchy-Keys are unique for each SMILES. This meant that the ID did not have to be customized. In addition, pickaxes were run on the COMMONS LAB server with the "enzymatic rule set" (250).

The reduction of the ruleset compared to the "Indermediat metacyc" or "generalised metacyc" allowed a faster runtime and the storage in a MongoDB (lotus_mines_enzymatic).

2.5 Dataprocessing

2.5.1 Part 1 : Datacleaning (only for working with files)

This step is necessary if the files have been split. This was the case for the prediction of the molecules on the IBU cluster, as it does not allow MongoDB (missing admin rights). At the beginning the file was split to run the prediction in parallel. Later this was replaced by a command in the terminal. This step is not required on the Commonslab server, as the files are entered directly into the MongoDB.

Before the data can be used as a database, first the possible duplicates has to be removed. This is the case because the files have been splitted. The first step is to adjust the ID for each file to ensure that there are no duplicates in the "ID" column.

Next, all compound files are read in and compared with each other. The binomial coefficient can be used to calculate the necessary iterations

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

In this case, k = 2 and n = 50, which leads to

$$\binom{50}{2} = 1225$$

comparisons. Compound file 1 is compared with compound file 2 first and all duplicates in compound file 2 are deleted. The deleted ID names are overwritten in reaction file 2 with the ID from compound file 1 so that the reaction is still correct. Reaction file 1 and reaction file 2 are compared again and the duplicates in reaction file 2 are deleted.

This step was not completed because we reduced the number of reactions (EnzymaticReactionRules.tsv - SMARTS) at an early stage. The prediction of the new molecules was then stored in MongoDB on the Commons-Lab server. As already mentioned, this practice should be avoided if possible. It is more error prone and time consuming.

More information about this scripts and and how to use it can be found on GitHub https://github.com/commons-research/read-MINE-results.

2.5.2 Part 2 : Database (MongoDB)

Efficient systems are needed to read in such large amounts of data and read it out in a reasonable time. Polars (python module) for example, can be used for this. It can read from various databases as well as *.parquet files.

Later during the master thesis (second part) the MongoDB was used, which saved the new components with Pickaxe directly in MongoDB. This saved the cleaning step and the data can also be easily indexed, which shortens the process time.
MongoDB Compass was then used to visualize the database. It also assist to build a “query” which can be translated to a python code in MongoDB compass itself. MongoDB Compass is a GUI that belongs to MongoDB. This program illustrates the database and provides assistance with filtering and indexing.

More information about the scripts and visualization of the database can be found on GitHub https://github.com/commons-research/expanded-naturalproduct-chemspace. The MongoDB can be accessed through the COMMONS server. The database name is “lotus_mines_enzymatic”. For access rights please contact Pierre-Marie Allard or by setting up their own server and run pickaxe with MongoDB. More information can be found here: https://github.com/commons-research/MINE-Database.

3 Results

3.1 Ruleset

The ruleset with the reactions to predict new compounds has been analyzed shortly, to get a better understanding of the matter at hand.

To get an overview about the most commen atoms used in the reactions, we looked for the reactants and the products. In the table 5 we can see the percentage over all enzymtic reactions (250).

atom	**reactants**	**products**
C	47.6 %	25.2 %
N	11.4 %	6.1 %
H	18.2 %	9.6 %
O	13.1 %	7.0 %
*	9.7 %	5.2 %

Table 5: atom distribution from ruleset "enzymatic"

In both cases, the atoms C, H, O and N are the most abundant. For the reactants they are describing 90.3% and for the products 47.9%.

3.2 MINES

The generated database "lotus_mines_enzymatic" is approx. 2.5 GB in size. It contains 3'432'312 compounds, of which 3'284'418 are predicted compounds and 147'861 are starting compounds. This starting compounds represents comounds without stereochemistry. For the reactions have been found 3'435'078.

The plots in Figure 7 show the 20 most frequently found compounds. The 12 most frequent compounds are found more than 1000x in all reactions. The most common starting compound is "alpha D-galactose", which occurs 15'552 (0.34 %) times in all reactions. It is followed by methanol (14'947), acetone (11'234), hydrochloric acid (10'060), formic acid (6'328) and methane (5'878). 362 starting compounds were found that only occurred once in all reactions.

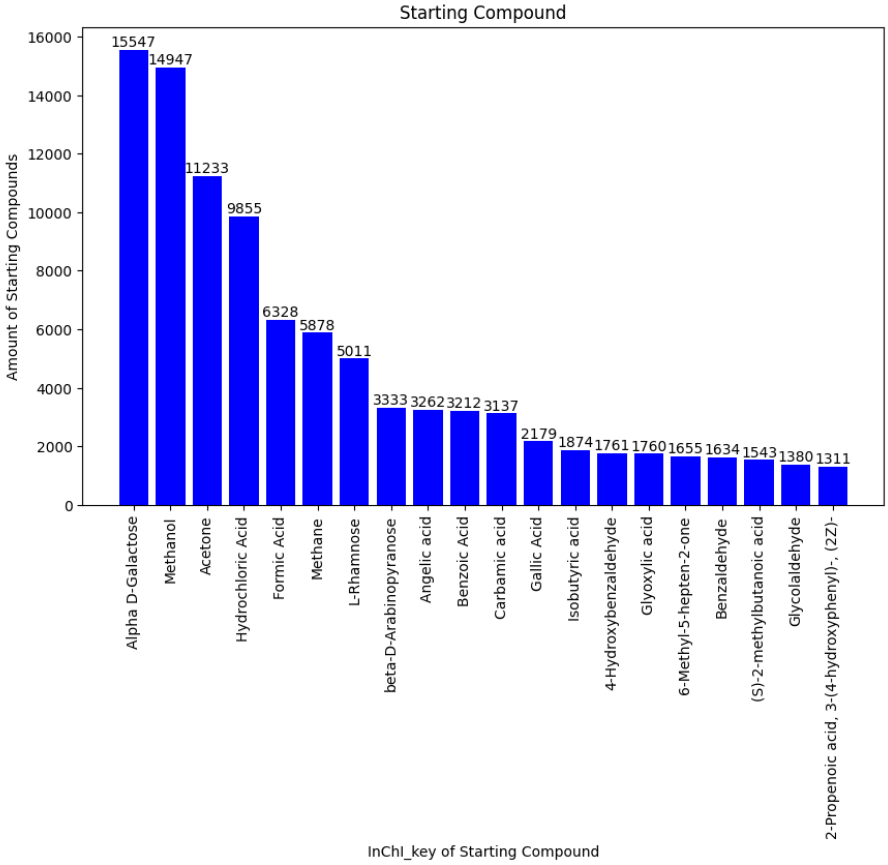

Figure 7: frequency of starting compounds

Brom (Br) is the most common of the predicted compounds (0.2 % with 6'931 compounds). The second most common molecule is "peroxymonosulfuric acid" (1'206), which, however, occurs more than 5 times less frequently than Br. Br is found in 28'341 reactions (0.8% of all reactions). The third most common element is

beta-D-apiose (636) followed by 2-hydroxy-2-methylpropanal (515). Of these, 3'213'470 predicted compounds occur only once in all of the reactions.

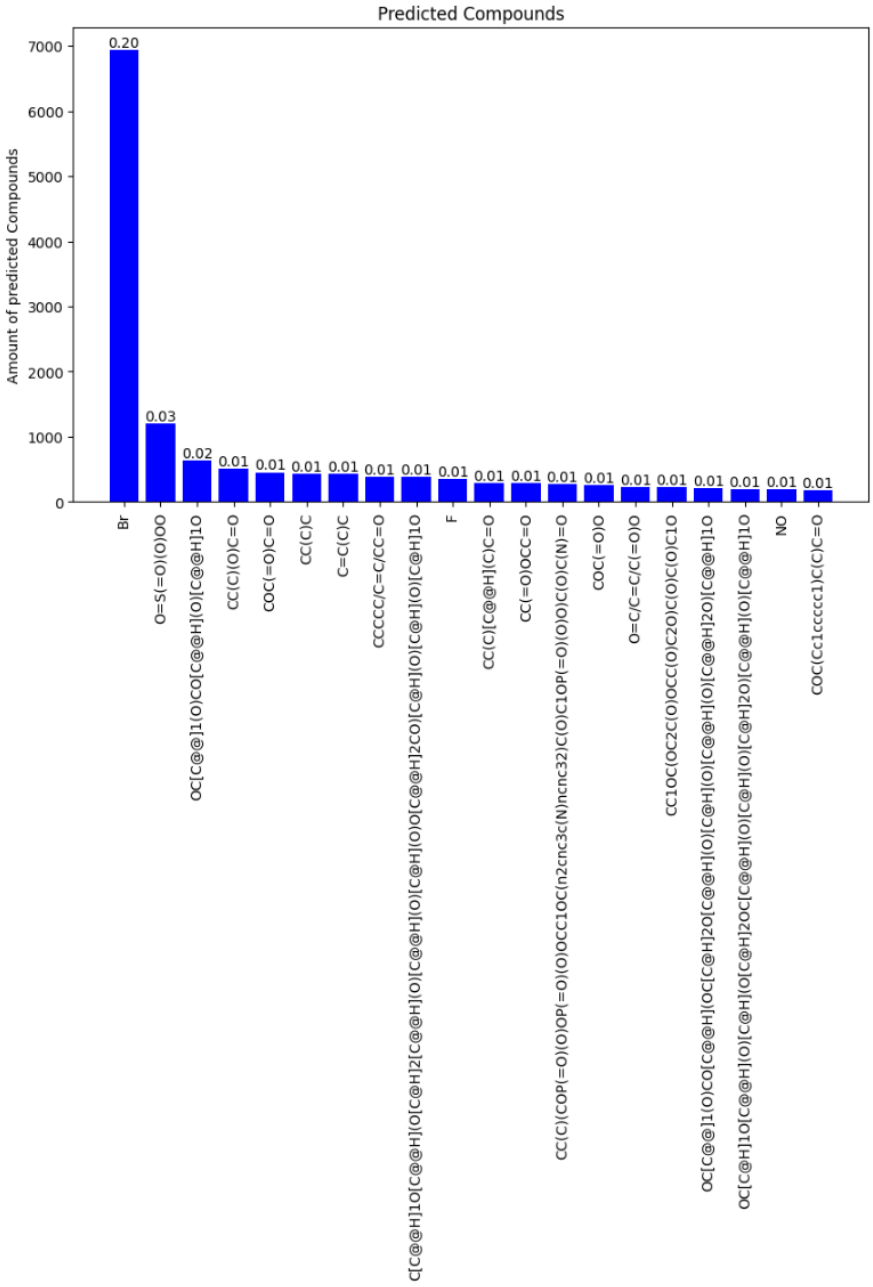

Figure 8: frequency of predicted compounds

To measure the discovering of the chemical space, we plotted some PCA to see the difference in the 3 most important parameters. From the 10 input compounds, 9 were loaded and labeled as "Starting Compounds". With the enzymatic rules (250) and the 33 "coreactants" it produced 109 "Predicted" compounds. The

chemspace (PCA) shows the distribution of the compounds.

The 2D PCA in figure 9 explains around 42% of the total variance. The 3d PCA in figure 10 explains around 50% of the total variance.

The PCA of the chemical space shows a broad distribution of the predicted compounds around the starting compounds, with the coreactant compounds only partially covered.

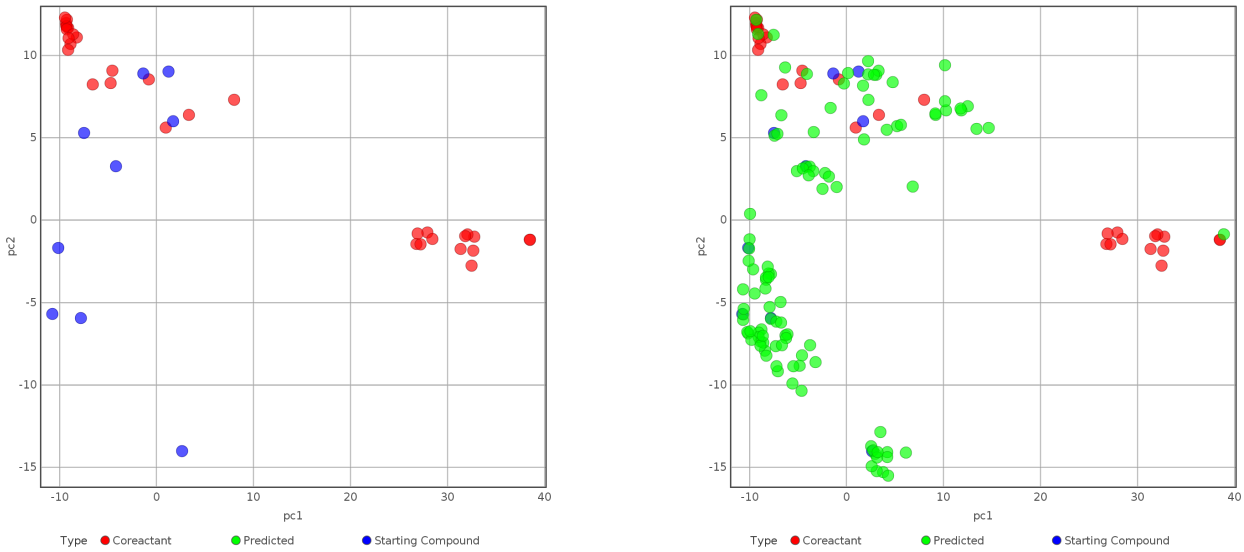

Figure 9: 2D PCA - (Left) starting compounds and coreactants; (Right) with predicted compounds

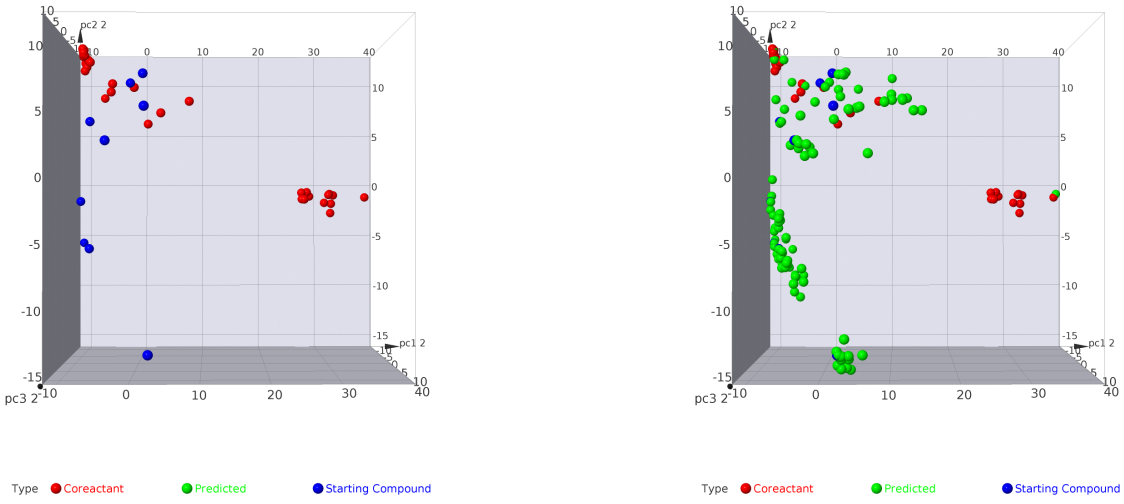

Figure 10: 3D PCA - (Left) starting compounds and coreactants; (Right) with predicted compounds

The explained variance can be seen in the table 6.

Principal Component	
PC1	29.383 %
PC2	13.048 %
PC3	9.368 %
PC4	5.061 %
PC5	4.453 %

Table 6: Explained variance percentage of the first 10 Principal Components

To make the link between the biological taxonomy and the expanded chemical structure, all the "Predicted Compounds" where annotated to the taxonomy of the starting compound. The figure 11 shows the change of the taxonomy from the starting compounds to the predicted compounds on the taxonomy "phylum". It shows an increase of 5.4 % in streptophyta. This belongs to the clade of plants.

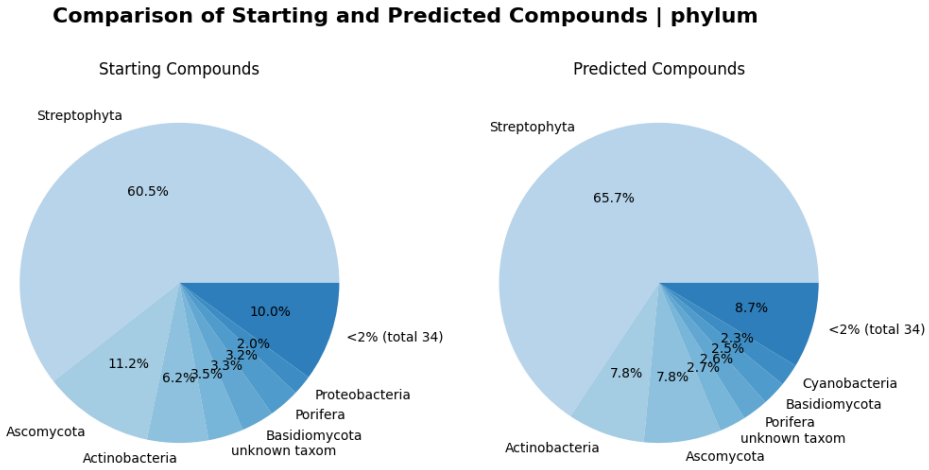

Figure 11: frequency of predicted compounds

The change in the expansion of the compounds and the change of the frequencies for each taxon can be seen in the tables below. When we compare the main expansion, which is the predicted compounds divided by the starting compounds, then the expansion is 23.2x. If we look at the different taxonomy, we can see a bigger expansion in the last column (~55x).

	Starting Compound		Predicted Compound			
domain	**counts**	**frequency**	**counts**	**frequency**	**difference**	**expansion**
Eukaryota	131741	87.35 %	7006150	85.62 %	−1.73 %	53.18 x
Bacteria	15566	10.32 %	1053386	12.87 %	2.55 %	67.67 x
unknown taxom	3467	2.30 %	121526	1.49 %	−0.81 %	35.05 x
<2% (total 1)	54	0.04 %	2045	0.02 %	−0.01 %	37.87 x

Table 7: domain taxonomy comparison

	Starting Compound		Predicted Compound			
kingdom	**counts**	**frequency**	**counts**	**frequency**	**difference**	**expansion**
Archaeplastida	96583	62.64 %	5448568	66.58 %	3.94 %	56.41 x
Fungi	22669	14.70 %	1289144	15.75 %	1.05 %	56.87 x
unknown taxom	21119	13.70 %	848839	10.37 %	−3.32 %	40.19 x
Metazoa	13807	8.96 %	596556	7.29 %	−1.67 %	43.21 x

Table 8: kingdom taxonomy comparison

	Starting Compound		Predicted Compound			
phylum	**counts**	**frequency**	**counts**	**frequency**	**difference**	**expansion**
Streptophyta	94949	60.54 %	5373456	65.67 %	5.12 %	56.59 x
Ascomycota	17552	11.19 %	640796	7.83 %	−3.36 %	36.51 x
Actinobacteria	9726	6.20 %	637600	7.79 %	1.59 %	65.56 x
unknown taxom	5557	3.54 %	222059	2.71 %	−0.83 %	39.96 x
Basidiomycota	5203	3.32 %	213996	2.62 %	−0.70 %	41.13 x
Porifera	4947	3.15 %	201234	2.46 %	−0.70 %	40.68 x
Proteobacteria	3145	2.01 %	186016	2.27 %	0.27 %	59.15 x
<2% (total 34)	15751	10.04 %	707950	8.65 %	−1.39 %	44.95 x

Table 9: phylum taxonomy comparison

	Starting Compound		Predicted Compound			
class	**counts**	**frequency**	**counts**	**frequency**	**difference**	**expansion**
Magnoliopsida	89927	56.57 %	5087111	62.17 %	5.60 %	56.57 x
unknown taxom	19183	12.07 %	1146891	14.02 %	1.95 %	59.79 x
Sordariomycetes	6775	4.26 %	252897	3.09 %	−1.17 %	37.33 x
Eurotiomycetes	6638	4.18 %	217101	2.65 %	−1.52 %	32.71 x
Agaricomycetes	5077	3.19 %	204650	2.50 %	−0.69 %	40.31 x
Demospongiae	4584	2.88 %	195820	2.39 %	−0.49 %	42.72 x
<2% (total 129)	26780	16.85 %	1078637	13.18 %	−3.67 %	40.28 x

Table 10: class taxonomy comparison

	Starting Compound		Predicted Compound			
order	**counts**	**frequency**	**counts**	**frequency**	**difference**	**expansion**
Asterales	16015	8.79 %	649343	7.94 %	−0.86 %	40.55 x
Lamiales	9437	5.18 %	525103	6.42 %	1.24 %	55.64 x
Fabales	8912	4.89 %	521638	6.37 %	1.48 %	58.53 x
Sapindales	8291	4.55 %	468314	5.72 %	1.17 %	56.48 x
unknown taxom	7820	4.29 %	399925	4.89 %	0.59 %	51.14 x
Streptomycetales	7641	4.19 %	363912	4.45 %	0.25 %	47.63 x
Malpighiales	6979	3.83 %	295734	3.61 %	−0.22 %	42.37 x
Eurotiales	6448	3.54 %	269201	3.29 %	−0.25 %	41.75 x
Gentianales	6259	3.44 %	224759	2.75 %	−0.69 %	35.91 x
Asparagales	4422	2.43 %	211735	2.59 %	0.16 %	47.88 x
Rosales	4084	2.24 %	211501	2.58 %	0.34 %	51.79 x
Apiales	3800	2.09 %	198911	2.43 %	0.34 %	52.35 x
Ranunculales	3705	2.03 %	179162	2.19 %	0.16 %	48.36 x
<2% (total 474)	88335	48.50 %	3663869	44.77 %	−3.72 %	41.48 x

Table 11: order taxonomy comparison

	Starting Compound		Predicted Compound			
family	**counts**	**frequency**	**counts**	**frequency**	**difference**	**expansion**
Asteraceae	15757	8.29 %	623998	7.63 %	−0.66 %	39.60 x
unknown taxom	8577	4.51 %	525103	6.42 %	1.91 %	61.22 x
Fabaceae	8400	4.42 %	443032	5.41 %	1.00 %	52.74 x
Streptomycetaceae	7641	4.02 %	441920	5.40 %	1.38 %	57.84 x
Trichocomaceae	6309	3.32 %	304668	3.72 %	0.40 %	48.29 x
Lamiaceae	6104	3.21 %	206562	2.52 %	−0.69 %	33.84 x
Rutaceae	3851	2.03 %	141686	1.73 %	−0.29 %	36.79 x
<2% (total 1506)	133466	70.21 %	5496138	67.16 %	−3.04 %	41.18 x

Table 12: family taxonomy comparison

	Starting Compound		Predicted Compound			
tribe	**counts**	**frequency**	**counts**	**frequency**	**difference**	**expansion**
unknown taxom	97860	55.22 %	5081667	62.10 %	6.88 %	51.93 x
<2% (total 655)	79367	44.78 %	3101440	37.90 %	−6.88 %	39.08 x

Table 13: tribe taxonomy comparison

	Starting Compound		Predicted Compound			
genus	**counts**	**frequency**	**counts**	**frequency**	**difference**	**expansion**
Streptomyces	7493	3.36 %	510760	6.24 %	2.88 %	68.16 x
unknown taxom	4752	2.13 %	170553	2.08 %	−0.05 %	35.89 x
<2% (total 7576)	210512	94.50 %	7501794	91.67 %	−2.83 %	35.64 x

Table 14: genus taxonomy comparison

	Starting Compound		Predicted Compound			
species	**counts**	**frequency**	**counts**	**frequency**	**difference**	**expansion**
unknown taxom	19150	6.88 %	728828	8.91 %	2.02 %	38.06 x
<2% (total 28168)	259011	93.12 %	7454279	91.09 %	−2.02 %	28.78 x

Table 15: species taxonomy comparison

	Starting Compound		Predicted Compound			
varietas	**counts**	**frequency**	**counts**	**frequency**	**difference**	**expansion**
unknown taxom	146759	99.08 %	8149800	99.59 %	0.51 %	55.53 x
<2% (total 199)	1362	0.92 %	33307	0.41 %	−0.51 %	24.45 x

Table 16: varietas taxonomy comparison

The 3'435'078 reactions describes mainly the full chemical transformation of the starting compounds with the coreactants to the predicted compounds. There are also more reactions then compounds found.

4 Discussion

At the beginning of the master's thesis, the scripts were mainly run on the IBU high-performance computer in Bern. However, I had no admin rights on the IBU cluster and no MongoDB was installed. This meant that I mainly worked with tsv (or parquet files). After the predicted compounds were saved in parquet files, I tried various queries with Polars (Python module). For example the "predicted compound" XY can have which taxonomy? For this, the "predicted compound" must first be found and then the possible starting compounds evaluated. Once the compounds are evaluated, the LOTUS database is used to search through all possible taxonomies with the given "starting compound".

The parquet files to be searched were over 100 Gb (> 1 billion compounds) in size and a simple query quickly took 20 minutes. This was unwieldy and was therefore not further elaborated.

Instead, we switched to the lab's own server, where MongoDB could be installed and the ruleset reduced. Instead of the "generalised" ruleset with 1024 reactions, the "enzymatic" ruleset with 250 reactions was used. This resulted in "only" around 3.5 million compounds, which was much faster for handle the data.

The "enzymatic" ruleset confirmed the biology based on the atom distribution. The most common atoms are carbon (C), hydrogen (H), oxygen (O) and nitrogen (N). This is also obvious, as the reactions mainly consist of enzyme reactions, which in biology are mainly responsible for the conversion of substances.

The most frequently represented "Starting Compounds" all seem to be represented with a relatively low percentage (< 0.5%). The most common "Starting Compound" D-Galactose is also known as brain sugar since it is a component of glycoproteins (oligosaccharide-protein compounds) found in nerve tissue. Galactose is found in most living organisms as a building block of oligo- and polycondensates of carbohydrates in various mucous membranes, from which the German name is derived ("Schleimzucker").

Boron (Br) as the most common "predicted compound" is not the first thing that comes to mind. We could

find some evidence, that it is an essential compound for the biology. Br^- is a required cofactor for peroxidasin, which catalyzes the formation of sulfilimine crosslinks. These posttranslational modifications are essential for tissue development and the structural integrity of the collagen IV scaffold within basement membranes (BMs). (McCall et al. 2014)

The chemical space shows mainly a broad distribution around the starting compounds. This is highly possible, because the "big molecules" are the "starting compounds" and with some small changes with the coreactants, it will stay around this molecules.

The taxonomy for the starting compounds compared to the predicted compound stays almost the same in the frequency. But the expanding rate is higher then from the original expanded molecules. Instead of 3'284'418 compounds, we have 8'183'107 predicted compounds of taxonomy, which describes the increase expanding rate of 55.34x. This is because more than one reaction can produce the same molecul. If we would filter for unique predicted compounds and taxonomy, the expanding rate would be at arround 25x, what we calculated in the beginning.

The next step from this thesis could be to produce spectras, which can be compared with the data of the massspectrometer. If this pipeline is established, the chemspace can be further expand with new rules to scale it up (for example "generalized ruleset"). With that, it can be eventual possible to find new compounds and also see a potential source (taxonomy) of it.

5 References

5.1 Abbreviations

MS : Massspectrometry
NP : Natural Product
InChI : International Chemical Identifier (represents chemical structures in a linear string)
InChIKey : hashed version of an International Chemical Identifier Key (unique identifier for one molecule and easy to search)
SMARTS : SMiles ARbitrary Target Specification
SMILES : Simplified Molecular Input Line Entry System (represents chemical structures in a linear string)

5.2 Bibliography

"All Natural." 2007. *Nature Chemical Biology* 3 (7): 351. https://doi.org/10.1038/nchembio0707-351.

Balietti, Stefano, Michael Mäs, and Dirk Helbing. 2015. "On Disciplinary Fragmentation and Scientific Progress." Edited by Daniele Fanelli. *PLOS ONE* 10 (3): e0118747. https://doi.org/10.1371/journal.pone.0118747.

Chainani, Yash, Geoffrey Bonnanzio, Keith Ej Tyo, and Linda J Broadbelt. 2023. "Coupling Chemistry and Biology for the Synthesis of Advanced Bioproducts." *Current Opinion in Biotechnology* 84 (December): 102992. https://doi.org/10.1016/j.copbio.2023.102992.

Ebbels, Timothy M. D., Justin J. J. Van Der Hooft, Haley Chatelaine, Corey Broeckling, Nicola Zamboni, Soha Hassoun, and Ewy A. Mathé. 2023. "Recent Advances in Mass Spectrometry-Based Computational Metabolomics." *Current Opinion in Chemical Biology* 74 (June): 102288. https://doi.org/10.1016/j.cbpa.2023.102288.

Jarmusch, Alan K., Mingxun Wang, Christine M. Aceves, Rohit S. Advani, Shaden Aguirre, Alexander A. Aksenov, Gajender Aleti, et al. 2020. "ReDU: A Framework to Find and Reanalyze Public Mass Spectrometry Data." *Nature Methods* 17 (9): 901–4. https://doi.org/10.1038/s41592-020-0916-7.

Jeffryes, James G, Ricardo L Colastani, Mona Elbadawi-Sidhu, Tobias Kind, Thomas D Niehaus, Linda J Broadbelt, Andrew D Hanson, Oliver Fiehn, Keith E J Tyo, and Christopher S Henry. 2015. "MINEs: Open Access Databases of Computationally Predicted Enzyme Promiscuity Products for Untargeted Metabolomics." *Journal of Cheminformatics* 7 (1): 44. https://doi.org/10.1186/s13321-015-0087-1.

Kim, Yeji, Jae Yong Ryu, Hyun Uk Kim, Woo Dae Jang, and Sang Yup Lee. 2021. "A Deep Learning Approach to Evaluate the Feasibility of Enzymatic Reactions Generated by Retrobiosynthesis." *Biotechnology Journal* 16 (5): 2000605. https://doi.org/10.1002/biot.202000605.

McCall, A. Scott, Christopher F. Cummings, Gautam Bhave, Roberto Vanacore, Andrea Page-McCaw, and Billy G. Hudson. 2014. "Bromine Is an Essential Trace Element for Assembly of Collagen IV Scaffolds in Tissue Development and Architecture." *Cell* 157 (6): 1380–92. https://doi.org/10.1016/j.cell.2014.05.009.

Mongia, Mihir, and Hosein Mohimani. 2021. "Repository Scale Classification and Decomposition of Tandem Mass Spectral Data." *Scientific Reports* 11 (1): 8314. https://doi.org/10.1038/s41598-021-87796-6.

Rutz, Adriano, Maria Sorokina, Jakub Galgonek, Daniel Mietchen, Egon Willighagen, Arnaud Gaudry, James G Graham, et al. 2022. "The LOTUS Initiative for Open Knowledge Management in Natural Products Research." *eLife* 11 (May): e70780. https://doi.org/10.7554/eLife.70780.

Shebek, Kevin M., Jonathan Strutz, Linda J. Broadbelt, and Keith E. J. Tyo. 2023. "Pickaxe: A Python Library for the Prediction of Novel Metabolic Reactions." *BMC Bioinformatics* 24 (1): 106. https://doi.org/10.1186/s12859-023-05149-8.

Contents

5.3 Figures

List of Figures

1 comprehensed output of mines 8
2 Input SMILES (3D - above) get flattened (down) through the "cleaning" process of pickaxe . 9
3 from the inputfile (above) to the stored "starting compound" (down) / change in InChIKey N->O 10
4 description for the reaction rule0402 11
5 Example of decarboxylation with THC-9 12
6 decarboxylation reaction for pickaxe 12
7 frequency of starting compounds 17
8 frequency of predicted compounds 18
9 2D PCA - (Left) starting compounds and coreactants; (Right) with predicted compounds . . 19
10 3D PCA - (Left) starting compounds and coreactants; (Right) with predicted compounds . . 19
11 frequency of predicted compounds 20

5.4 Tables

List of Tables

1 LOTUS dataframe columns - chemical structure 4
2 LOTUS dataframe columns - organism structure 5
3 LOTUS dataframe columns - reference 5
4 change in the InChIKey N->O 9
5 atom distribution from ruleset "enzymatic" 16
6 Explained variance percentage of the first 10 Principal Components 20

7 domain taxonomy comparison 21

8 kingdom taxonomy comparison 21

9 phylum taxonomy comparison 21

10 class taxonomy comparison 22

11 order taxonomy comparison 22

12 family taxonomy comparison 23

13 tribe taxonomy comparison 23

14 genus taxonomy comparison 23

15 species taxonomy comparison 23

16 varietas taxonomy comparison 24

5.5 Links

All the scripts can be found unter the COMMONS Lab github website: https://github.com/commons-research. Diary and thoughts can be found on the dendron page: https://commons-research.github.io/commons-dws-public/notes/rf0mbo9ji3lqry16qmur71g/.

Appendix

Declaration of Authenticity

I declare that I completed the report independently and used only these materials that are listed. All materials used, from published as well as unpublished sources, whether directly quoted or paraphrased, are duly reported. I declare that the full report was written by me (Pascal Amrein).