

Chemoinformatical Analysis of Natural Products

Dissertation

zur Erlangung des Doktorgrades der Naturwissenschaften

(Dr. rer. nat.)

der

Naturwissenschaftlichen Fakultät II

Chemie, Physik und Mathematik

der Martin-Luther-Universität Halle-Wittenberg

vorgelegt von

Frau Anne-Kathrin Hartig

1. Gutachter: Prof. Dr. Ludger A. Wessjohann

2. Gutachter: Prof. Dr. Christoph Steinbeck

Datum der öffentlichen Verteidigung: 26.09.2024

Danke!

Mein erster Dank geht an Prof. Dr. Wessjohann, der mir die Möglichkeit geboten hat, an diesem spannenden cheminformatischen Thema zu arbeiten und zu wachsen. Ich möchte Ihnen für die Freiheiten danken die Richtung der Arbeit mit lenken zu können, aber auch für die vielen hilfreichen Denkanstöße! Ohne PD Dr. Brandt, der immer vor Ort war und ein offenes Ohr hatte, wäre die Umsetzung so aber nie möglich gewesen – vielen Dank Wolfgang! Genauso gilt mein Dank den anderen Mitglieder des Projektes Prof. Dr. Müllner-Riehl, Dr. Schnitzler, Dr. Franke, Annegret und Laura für die produktiven Diskussionen. Ohne die wundervolle Arbeitsgruppe Computerchemie am IPB hätte die Arbeit nur halb so viel Spaß gemacht. Danke Daniela, Silke, Evelyn, Marius, Anne, Jördis, Richard und Mathias für das entspannte Arbeitsklima, die Freundschaft, aber auch die anregenden Diskussionen. Auch im privaten Umfeld wurde ich immer unterstützt – Danke Mama, Papa, Caro und Marcus, dass ihr immer für mich da seid und an mich glaubt!

Contents

Abstract	I
List of Figures	III
List of Tables	IV
List of Abbreviations	V
1 Introduction	1
1.1 Natural Products and their Natural Sources	1
1.2 Text Mining	3
1.3 Scope of the Thesis	5
2 Secondary Metabolites of Indonesian Plants	8
2.1 Material and Methods	9
2.2 Analyses	11
2.3 Results and Discussion	15
2.4 Summary	23
3 Creation of a Species - Natural Product Gold Set	24
3.1 Material and Methods	25
3.2 Results and Discussion	32
3.3 Summary	36
4 Creation of a Global Data Set	38
4.1 Material and Methods	39
4.2 Results and Discussion	53
4.3 Summary	60
5 A World Wide Overview of Plant Natural Products	62
5.1 Material and Methods	63
5.2 Results and Discussion	64
5.3 Summary	72
6 Discussion and Summary	73

Table of Contents

References	i
Appendix	xv

Abstract

The importance of bioactive natural products from plants has been known in medicine for thousands of years. Even in modern medicine, many drugs are still based at least on the chemical structure of a natural products. Due to the progressing resistances it is of high importance to find new anti-infective substances. Often, in plant families where bioactive substances have already been found, further potentially active natural substances are searched for. Other research groups focus on specific natural product families. In this work, a more comprehensive approach is presented that links bioactivities to the entire plant phylogeny, not just that within a family, and incorporates spatial data.

First, this approach was applied to a limited area. For this purpose, the island of Java was selected as one of the biodiversity hotspots. Which plants occur in this region was determined once by the Flora Malesiana, as well as the Flora of Java, supplemented by concrete GBIF-findings of plants. A phylogentic tree was generated for these plants, and information on natural products and their bioactivities was collected from databases. Combining the phylogeny with the bioactivity information can reveal plant families that are proportionally less studied than others and thus contain unknown natural products. But it can also show for which families significantly more or less bioactivities have been detected compared to the others. The analysis of spatial data, on the other hand, can show in which regions plants were collected more frequently or less frequently, but one can also model biodiversity. Using the biodiversity, one can also model a chemodiversity with the information of the detected natural products for the individual species. In the case of Java, the available data show a linear correlation of chemical and biological diversity.

The approach tested for Java should then be applied to a broader, global data set. To achieve the most complete data possible, the primary literature should be used as the source. To be able to cope with this, a text mining approach was chosen. To develop an appropriate tool, a gold standard corpus of 100 publications was first created manually. In this, all relations of biological sources and natural products extracted from them were marked. This was divided into a test and a training set for the development of the text mining tool. The tool includes first the recognition of biological sources and natural products (Named Entity Recognition), which is based on rules but also on already available trained models. Furthermore, it implies a rule-based relation extraction, i.e. the extraction of biological source - natural product pairs. This extraction is not

only performed in continuous text, but also in tables.

For the natural products found, chemical structures as well as activity data were subsequently added in databases. The resulting data set was analyzed comprehensively. On the one hand, the natural products were analyzed with respect to their chemical classes, and on the other hand, they were again enriched with phylogenetic and spatial data. It was also shown that the dataset can be valuable for structure-activity relationship analyses.

List of Figures

1.1	Native species richness	6
2.1	Habitats of Indonesia	17
2.2	Bio- and Chemodiversity of Java.	19
2.3	Structural classes of natural products from Java.	21
2.4	Taxonomic analysis of antiinfective effects in seed plant species across the flora of Java.	22
4.1	Scheme of text mining process	40
4.2	Scheme of tables searched for relations	47
4.3	Database diagramm of the resulting database.	50
4.4	Number of Publications used for the text mining per year of publication.	54
4.5	Number of natural products per kingdom	55
4.6	Ratio of chemical substructures of natural products per kingdom	57
4.7	Venn diagramm of activity predictions	59
5.1	Phylogentic tree of extracted plants.	66
5.2	Global bio- and chemodiversity	69
5.3	Distribution of plants per habitat	71
A1	Visualization of frequency of occurance of chemical classes in natural sources	xli
A2	World map with ratio of natural product containing plants.	xlii

List of Tables

2.1	Comparison of KNApSAcK and DNP	16
3.1	Number of entity annotations	33
3.2	Distribution of annotated “species produces natural product” relations	35
4.1	Simplified regular expressions for the relation detection for “species produces natural product”-relations	46
4.2	Natural Product Matching Resources	51
4.3	Precision, Recall and F-Score	53
4.4	Number of text bases	55
4.5	Number of unique natural products	56
4.6	Number of unique natural products	58
A1	KEGG metabolic pathways included in the selection of secondary metabolites	xv
A2	SMARTS descriptors for natural product classes	xxi
A3	Percentage of chemical classes per kingdom	xxvi
A4	Taxonomic analysis of anti-infective effects.	xxvii
A5	Environmental variables used for the species distribution modelling	xxx
A6	Trigger and negation words	xxx
A7	ChEMBL Bioassays	xxxv
A8	ChEMBL targets	xxxvi
A9	ChemOnt classes	xxxvi

List of Abbreviations

Abbreviation	Full Name
DNP	Dictionary of Natural Products
TCMD	Traditional Chinese Medicine Database
NPASS	Natural Product Activity & Species Source Database
KEGG	Kyoto Encyclopedia of Genes and Genomes
MOE	Molecular Operating Environment
GBIF	Global Biodiversity Information Facility
WWF	World Wide Fund For Nature
API	Application Programming Interface
TSS	True Skill Statistic
EBDCS	Economic Botany Collection Standard
CI	Confidence interval
IUPAC	International Union of Pure and Applied Chemistry
OCID	Ontology Concept ID
NER	Named Entity Recognition
ML	Machine Learning
NLTK	Natural Language ToolKit
SMARTS	SMILES Arbitrary Target Specification
InChI	International Chemical Identifier
HTS	High-Throughput Screening
SAR	Structure Activity Relationship
RF	Random Forest
SVM	Support Vector Machine
DNN	Deep Neural Network
PLS-DA	Partial Least Square, Modified to a Discriminant Variant
AI	Artificial Intelligence

1 Introduction

1.1	Natural Products and their Natural Sources	1
1.2	Text Mining	3
1.3	Scope of the Thesis	5

1.1 Natural Products and their Natural Sources

“The simplest definition for a natural product is a small molecule that is produced by a biological source.”¹ Those biological sources can be plants, fungi, algae, microbes or even mammals or insects. However, most often the term natural products describes secondary metabolites. Those organisms produce the organic compounds for protection, competition or inter- and intra-species interactions. Unlike primary metabolites they are not necessarily for growth, development or reproduction of the organisms. In the following natural products are defined as small molecules produced by plants, fungi, algae or microbes as secondary metabolites. Therefore, DNA, RNA, proteins and other polymers are not taken into consideration.

Humans use nature since millenia in their traditional medicines. Starting with the use of *Papaver somniferum* (poppy) as early as 4000 B.C. as an analgesic due to the main alkaloids morphine and codeine², natural products were the main source of new drugs until the 1980s. Since then, it appears that combinatorial chemistry has taken the lead in drug discovery. However, Newman *et al.* have shown that despite this, 49 % of all drugs approved between 1981 and 2014 are either natural products or directly derived from them. Only 22 to 33 % of the antiinfective drugs approved in this period have active ingredients of synthetic origin, while the vast majority are natural products, or their derivatives or mimetics.³

Scaffolds (the core molecular structure common to each class) of natural products are often used as basis for the further development and improvement of drugs. Antibiotics are a prominent example. Nine antibiotic classes, which can be differentiated based on their scaffolds, contribute to most of the clinically approved antibiotics today. The incremental modification of available natural scaffolds has become the prevailing approach of antibiotic drug discovery.^{3,4} Most of these were developed between the mid-1930s and early 1960s, and only three new classes were introduced after an innovation gap of around 40 years⁴, despite having already been known for at least two decades. While combinatorial chemistry has allowed the pharmaceutical industry to produce and screen vast numbers of molecules, this approach was only moderately successful in finding compounds with new modes of action.^{3,5} Compound libraries produced through combinatorial chemistry might have larger numbers of compounds, but unless focused around an already known lead compound (new chemical entity with the potential to develop a new drug) they contain a lower rate of biologically relevant ones.²

Consequently, scientists are highly interested in natural products. However, in terms of natural product discovery most of them are concentrated on either some species, genera or one special structural type of natural products.⁶⁻⁸ Therefore it is not surprising that only a small part of the scientifically described plant species ($\sim 370,000$ ⁹) has been studied in terms of natural products.¹⁰ Many scientists expect the number of phytochemicals exceeding 500,000 based on the more than 200,000 which are already known.¹¹⁻¹⁴ A comprehensive overview which natural products are already described in which natural sources combined with bioactivity information can help to select target species for phytochemical analyses. Those information can be combined with phylogenetic information to predict chemical diversity¹⁵, white spots in the phylogenetic tree in terms of phytochemically studied plants, but can also serve to see which plant families were disproportionately examined. This analyses bases on the idea, that plants in their evolutionary history had to adapt to their environment and pathogens leading to a wide range of chemical compounds.¹⁶ Consequently chemical compounds will be shared among closely related ones.^{14,17,18} For example Zhu *et al.*¹⁹ found 17 “drug-producing” clusters across plants, bacteria, fungi and metazoa. 30 % of novel secondary natural products published between 2001 to 2011 were found in families outside these clusters.¹⁹ However, metabolic diversity may be limited by the scaffold producing pathways and therefore biological and chemical diversity might not correlate linearly.^{8,20}

Although there are some database, which can serve as comprehensive data collection,

such as the Dictionary of Natural Products (DNP)²¹, Reaxys²², TCMD (traditional Chinese medicine database)²³ or the KNApSACk^{24,25} database containing information of species and the natural products isolated from them, but the databases are either commercial, focusing again just on one part of the whole information available or are not comprehensive.²⁶ Recently Zeng *et al.* published a new open access database containing natural products and their respective sources called “Natural Product Activity & Species Source Database (NPASS)”. This database contains 25,041 source organisms consisting of 16,581 plant, 1,675 bacteria, 2,503 metazoa and 2,107 fungi species, with 35,032 corresponding natural products. For those 5,863 biological targets and 446,552 activity records are stored in NPASS database.^{27,28} Therefore this is the most comprehensive and freely available database concerning natural products and their sources, to our knowledge. Of course there are freely available databases containing a higher number of natural products (e.g. COCONUT²⁹), however those lack the information of the natural source. As described by Sorokina and Steinbeck²⁹ one main drawback of most available databases is the fact that they are not supported continuously.

1.2 Text Mining

Text mining approaches are increasingly relevant in natural sciences not only to retrieve more or better information faster, but to gain new knowledge from existing data. In life sciences, and especially in the biomedical domain they are used to detect a variety of bio-entities such as genes, proteins, diseases and chemicals, connected to genomics/transcriptomics, proteomics and metabolomics. Also, relations like protein-protein interactions, gene-disease relations or drug-drug interactions have been extracted automatically from texts.³⁰⁻³⁹ In order to train and test the reliability and detection rate of automated entity recognition or information extraction workflows, a set of documents is required where the entities and relations of interest are already annotated. These can be produced either manually or automatically for example by using dictionaries or some pre-trained annotation tools. The latter ones are called silver standard corpora, while the expert generated manual corpora are known as gold standard corpora since it is assumed that those annotations are the most complete and correct ones. In 2010 Blake found that typically only around 8 % of the scientific claims of a scientific article can be found in its abstract and stated the urgent need of text mining systems and corpora covering full text articles for training those systems.⁴⁰ This finding was recently confirmed by Westergaard *et al.* who compared 15 million full-text articles for protein-

protein, disease–gene, and protein subcellular associations with their corresponding abstracts.⁴¹ Reiterating the same, the differences in structure and content of abstracts versus the full text articles make it necessary to create corpora built on full text publications.^{35,41} The problem with abstracts is not only the limited comprehensiveness of information due to the required shortness, but also the selection bias, representing the authors' choice for a focus, or the scientific trend at the time it was published. Despite these limitations, there are several gold standard corpora available that are based only on abstracts or on single sentences in the biomedical domain. These are e.g. the CHEMDNER corpus for chemical entities⁴², the GENIA corpus with annotations from the biological domain⁴³ or BioInfer focusing on protein-protein interactions⁴⁴. However, some corpora include the full text, for example the CRAFT corpus consists of 97 full text articles annotated semantically as well as syntactically in the biomedical domain or the BioC-BioGrid corpus including 120 full text articles annotated for protein-protein and genetic interactions.⁴⁵ Regardless whether the corpora are based on full texts, abstracts or single sentences, to our knowledge there is until now no gold or even only a silver standard corpus available where annotated named entities and their relations are spanning over sentences, paragraphs or tables, although the corpus is based on full texts. Each gold or silver standard corpus is usually selected and designed with the text mining application to be trained in mind. These designs differ in the selection of documents, as well as the entities and eventually relations that are annotated. Therefore, such corpora are often of limited use to train text mining applications with another scope of interest. For example, if the corpus is based on documents which were selected for dealing with proteins it is probably not applicable for training recognition of cancer treatments.³⁵

There are two gold standard corpora dealing with the relationships between natural products (secondary metabolites) produced by species (plants, fungi, microbes or animals). Additional to this also two text mining tools exist. Jensen *et al.* trained a Naïve Bayes classifier to extract plant – phytochemical relationships from 21 million abstracts in PubMed/MEDLINE. They extracted 23,137 compounds from 15,722 plants through 369,549 edges. However, they did not publish a corpus for training and testing their Naïve Bayes classifier.⁴⁶ Choi *et al.* did publish a corpus consisting of 377 sentences from 245 PubMed abstracts. Here, 267 plant entities and 475 chemical entities are annotated as well as 1,007 chemical-plant relationships (550 positives and 457 negatives). Based on this corpus Choi *et al.* developed a rule based extraction model and tested it on a pre-annotated corpus yielding overall F-Scores of 68.0 % and 61.8 %.^{31,47,48} The

second corpus is the recently published ChEBI corpus.⁴⁹ This corpus consists of 200 abstracts and 100 full text publications. Additional to the “metabolite” and “species” entities the authors annotated chemicals, proteins, biological activity and spectral data. The relations “associated with”, “binds with”, “metabolite of” were marked beside “isolated from”. On average, the ChEBI corpus contains 48 metabolite and 33 species entity annotations together with 8 “isolated from”-relations per full paper.

Text mining approaches have been used successfully in the biomedical domain for detecting several bio-entities such as proteins, genes, diseases and chemicals. Also, relations like protein-protein interactions, gene-disease relations or drug-drug-interactions have been extracted automatically from texts.³⁰⁻³⁹

In addition to classic NLP methods, artificial intelligence is also increasingly used to train models that can extract information. The best known at the moment, ChatGPT⁵², you can ask questions based on a chatbot and receive a fully formulated answer. However, if you ask chatgpt for all known natural substances, you get the answer that chatgpt cannot enumerate them, but the best known databases (e.g. COCONUT) are mentioned (ChatGPT was accessed on 4/29/2023).

1.3 Scope of the Thesis

In this thesis, an approach to streamline and target the search for new antiinfective compounds is presented that integrates phylogenetic and spatial data with information on bioactivity first. As target country Indonesia was chosen as model region and plants as model organisms, since Indonesia is one of the 17 megadiverse countries⁵³ with more than 40,000 species of vascular plants in the Malesian region (see Figure 1.1). Furthermore, the Indonesian traditional medicine Jamu uses a vast amount of indigenous plants and there are two databases (KNApSAcK^{24,25} and NPASS^{27,28}) with natural products of plants in this region available. For the island of Java we had detailed information of indigenous and invasive plant species and combined taxonomic, phylogenetic, spatial and phytochemical information. By this approach we identified over-/underrepresentation of antiinfective activities across the phylogenetic tree and looked into the correlation of biodiversity and chemical diversity.

Eventually, the creation of a gold standard corpus is described. After investigating publications dealing with the isolation of natural products from species it became obvious, that many if not most relations span over more than one sentence or even

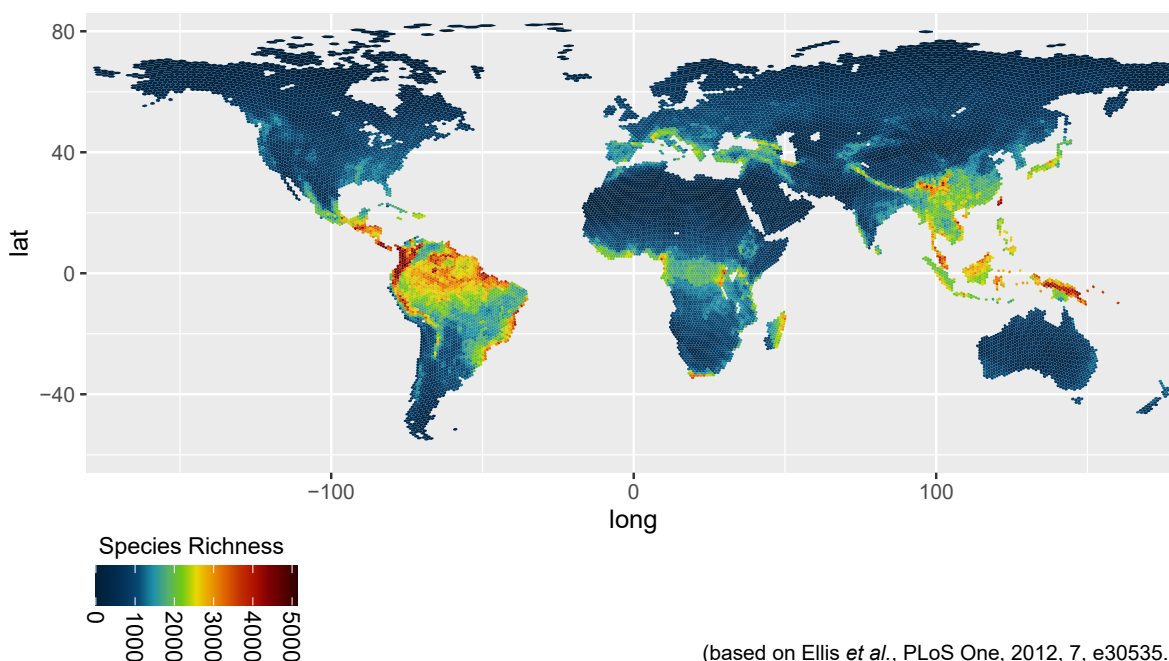


Figure 1.1: Native species richness based on Ellis *et al.*⁵⁰. Red indicates high native species richness, while blue represents low richness. One of the hotspots is located in and around Indonesia. (This picture was created using ggmap⁵¹)

paragraph. Additionally, information commonly is summarized in tables, partially with co-references to the running text. Therefore, we developed a gold standard corpus containing also relations spanning over more than one sentence or paragraph and including relevant information localized in tables. To our best knowledge, this it is the only gold standard corpus annotating relations spanning over multiple sentences, occurring in tables and including their co-references. This corpus consists of 102 freely available full text PubMed Central articles, where all species and chemicals are annotated as such together with all “species – produces – natural product(s)” relations. Based on this gold standard corpus a text mining tool was developed to create a comprehensive overview of all known natural products. As mentioned above the only comprehensive source of all available information about natural products and their source are full text publications. A short search in the PubMed Central database for natural product related publications (using several search terms such as phytochemistry, natural product or natural compound) ended up with about 75.000 results as for 2017-05-08.

The resulting database of the text mining approach is used to perform analysis similar to

those performed for the metabolites and plant species of Java along to chemoinformatic analysis. Furthermore, statistical analyses were used to prove that an unbiased collection of natural product - natural sources relations have been created.

2 Secondary Metabolites of Indonesian Plants

Parts of this chapter were already published in⁵⁴.

2.1	Material and Methods	9
2.1.1	Taxonomic and metabolite information	9
2.1.2	Phylogenetic Tree	11
2.2	Analyses	11
2.2.1	Natural product classes	11
2.2.2	Statistical analyses	11
2.2.3	Phylogenetic patterns	12
2.2.4	Bio- and Chemodiversity of Java	13
2.2.5	Habitats	15
2.3	Results and Discussion	15
2.3.1	Database crawling	15
2.3.2	Chemical Diversity	16
2.3.3	Metabolite information	20
2.3.4	Phylogenetic Diversity and Taxonomic Analysis	20
2.4	Summary	23

To see whether the approach to combine phylogenetic, geographic and natural product information of plants to gain a comprehensive overview and streamline the process of identifying potential new bioactive compounds, Indonesia and especially Java was chosen as a proof of concept before the approach is transferred to global application. Indonesia was chosen since it is one of the biodiversity hotspots.⁵³ Only a few plant species have been thoroughly screened for natural products and bioactives, therefore

regions with high biodiversity can provide more screened plants than other regions. Java was chosen from the more than 13,000 Indonesian islands because here the Flora of Java⁵⁵, a collection of all plants known to grow on Java, is available.

2.1 Material and Methods

2.1.1 Taxonomic and metabolite information

As starting point for the database queries, names of plant species growing on Java were used. This taxonomic information was extracted from the Flora of Java Vol. I-III⁵⁵, a systematic account of all seed plants registered on the island of Java. Additional taxa, described for Java between 1960 and 2014, and not included in the Flora of Java, were added from the major collections of the region, e.g. the Herbarium Naturalis (L, WAG & U), the herbarium at the Royal Botanic Gardens, Kew (K), the Australian National Herbarium (CANB), and the herbarium at the Smithsonian Institution (US). Furthermore, regional checklists (e.g.⁵⁶) were used as resource for the species names. All seed plants described as introduced in the Flora of Java⁵⁵ were marked as such. Using the Taxonomic Name Resolution Service^{57,58} all names were checked and if necessary corrected. In total the database queries started with 7,573 plant species of which 5,392 are native to Java (2,352 genera in total, 1,652 native).

The KNApSAcK^{24,25} and NPASS²⁷ databases contain information about species and metabolites produced by those, together with (bio-)activity information for the metabolites. Those two were chosen, since they are freely accessible and in case of the KNApSAcK, it also contains special information about the herbs used in traditional Indonesian medicine (Jamu). The databases were queried automatically using python scripts.

KNApSAcK provides an url that can be used to be incorporated into programs: [http://www.knapsackfamily.com/knapsack_core/info.php?sname=\[item\]&word=\[keyword\]](http://www.knapsackfamily.com/knapsack_core/info.php?sname=[item]&word=[keyword]) Item must be one of “organism”, “metabolite”, “formula”, “C_ID” or “CAS_ID” depending on what type your keyword is. In our case the item was organism, while the keyword was replaced by a species name. Using the python package requests⁵⁹ the http query was sent to the web and the response of the server can be read and processed. The processing of the returned html website was performed using the python package BeautifulSoup⁶⁰ and all desired information (names, identifier and structures of the metabolites connected to the species in this database) is stored in text files. An unique

set of the metabolite identifier (C_IDs) was used to query for metabolite activity. Here the python package mechanize⁶¹ was used, since the query page is a JavaServer page and it was necessary to click the button “C_ID” and fill an entry field with the metabolite C_ID. Therefore, it was not sufficient like in the first query case to manipulate an url. The response was parsed using regular expression to extract the activities of the metabolites. KNApSAcK was accessed 18 March - 5 April, 2016.

The NPASS queries in contrast are processed using JavaScript, for which it was necessary to use the python package selenium⁶². The workflow, however was similar: starting with the species names NPASS was queried for metabolite names, identifier (Natural Product ID) and structure. Here it was possible to follow the links from the resulting page to additional metabolite information where also the activities are listed. All resulting pages were parsed using the BeautifulSoup package to extract the information. NPASS was accessed 18-25 January, 2018.

From KNApSAcK 10,612 metabolites from 1,883 species and from NPASS 9,894 metabolites from 1,889 species were extracted, with an intercept of 1,776 species. The databases provided the names, in some cases with synonyms, and structures along with additional information, such as molecular formula or CAS number. The structures were available as InChi, InChiKey and SMILES string and in the KNApSAcK database in some cases as mol2 file. To identify duplicate structures, within one database, but also the intercept between both databases, the metabolites were normalized using their names and structures. Therefore, all metabolites with the respective identifier were loaded into a MOE (Molecular Operating Environment⁶³) database (mdb) and therein duplicate structures were identified and merged. In total, the queries resulted in 17,117 unique structures.

To decide whether a metabolite is primary or secondary the KEGG (Kyoto Encyclopedia of Genes and Genomes)⁶⁴⁻⁶⁶ database of metabolic pathways was used. Therefore, the python packages requests⁶⁷ and the module KEGG of the biopython package⁶⁸ were used. First using requests the KEGG API (Application Programming Interface) in URL form was queried with the metabolite names to check if KEGG contains pathway information for this compound. If there was no information available, it was assumed that the metabolite is secondary. If there was compound information, the respective identifier was extracted using regular expressions to query via the KEGG module of biopython in which pathways the metabolite is annotated. The KEGG module also uses the KEGG API, however, it already parses the returned html string and directly

returns all pathways in which the compound is mentioned. For example metabolites taking part in pathways from the photosynthesis, carbon metabolism, citrate cycle or fatty acid biosynthesis domains, as well as those taking part in hormone or nucleic acid synthesis were defined as primary a detailed list of all pathways defining metabolites as primary can be found in Table A1. 16,503 metabolites are defined as secondary metabolites in this dataset.

The associated biological activities were obtained from KNApSAcK and NPASS. In the following analyses, special attention is put on antiinfective activities, including anti-fungal, anti-bacterial, anti-parasitic and anti-viral (3,705 metabolites in total) and in detail only those affecting vertebrate pathogens (2,859 metabolites). The 2,859 secondary metabolites are connected to 1,640 species (941 genera), of which 796 are native (501 genera).

2.1.2 Phylogenetic Tree

Hinchliff & Smith⁶⁹ in 2014 published a phylogenetic tree consisting of 13,093 land plant genera, which corresponds to estimated 80 % of the existing plant diversity and 83 % of recognized land plant genera collected from GenBank⁷⁰. Puned versions of the tree with all genera not present on Java excluded was created for the phylogenetic analysis of the Java dataset. Two trees were constructed, built from the 1,426 native genera and from 2,083 genera (introduced and native plants), respectively. Therefore, 14 % (226 native genera) and 11 % (269 genera) of the genera are missing in the trees.

2.2 Analyses

2.2.1 Natural product classes

To classify the natural products to the different chemical classes SMART strings describing those classes were defined in Molecular Operating Environment (MOE). All SMARTS are shown along with additional requirements to be categorized into the respective class in Table A2.

2.2.2 Statistical analyses

For the analyses the relation and additional information obtained from the KNApSAcK databases, GBIF⁷¹ (Global Biodiversity Information Facility), WWF (World Wide Fund For Nature) terrestrial ecoregions⁷² information and the tree of life^{73,74} as taxonomy

were used. Testing the hypothesis that all plant families have the same ratio of active species, the confidence intervals of the ratio of active species per family in dependence to the number of species per family can be calculated using the following formula because the binominal distribution can be approximated by a normal distribution (central limit theorem):

$$x = p \pm z(1 - \alpha) * \sqrt{\frac{p * (1 - p)}{n}} \quad (1)$$

With p as the calculated mean of active species per family, $1 - \alpha$ 0.95 respectively 0.997 and n as number of species per family. With this calculated for all possible number of species per family one can then reject the hypothesis of having the mean ratio of active species if the ratio of the corresponding family is outside the confidence intervals with a probability of 95 % respectively 99.7 %, which does not automatically mean that they are significantly different. Other analyses performed were network analyses, where all families were connected which share metabolites, and statistics of the species in dependence on the habitats they are growing in. Furthermore, distribution plots of the species in Indonesia were created.

2.2.3 Phylogenetic patterns

In a first step it was checked if an underlying phylogenetic patterns in the distribution of antiinfective bioactivity exist. For this, the D statistics⁷⁵ were calculated using the implementation in the R package “caper”⁷⁶. D is a measurement to evaluate the phylogenetic signal of a binary characteristic, in this case the antiinfective activities. Therefore, it calculates the sum of the differences between sister and gap in the phylogenetic tree. This value is then compared with a clumped and a random distribution of the characteristic. These serve as references for a conserved evolutionary model approximated by Brownian motion^{75,77} and a trait randomly distributed over the phylogenetic tree for the given phylogeny and given prevalence of the characteristic (proportion of peaks in character state 1).

To identify clades in the phylogenetic tree which show a significantly higher or lower proportion of antiinfective metabolites, a phylogenetic clustering approach was used. Using the program phylocom⁷⁸ the algorithm “nodesig” is used. It was originally developed as metric of phylogenetic community structure. However, here the information “antiinfective bioactivity” yes or no was used as “communities”. The algorithm compares

randomly distributed bioactivity with the original distribution in several runs. With a significance level of 5 % clades are then chosen with over or under-represented bioactivity and classified as “hot” or “cold” clades.

2.2.4 Bio- and Chemodiversity of Java

One analysis aim was to evaluate the spatial patterns of the plants growing on Java to have an insight into the bio- and chemodiversity of the island. Therefore, plant occurrences were downloaded from the Global Biodiversity Information Facility⁷⁹ database for Indonesia (accessed 03 May 2018). Since for a large number of accessions no coordinates were given, but a description of the locality where the plant was observed, an approach developed by Gratton *et al.*⁸⁰ was applied. Thereby the locality is automatically georeferenced using coordinates of administrative units and/or geographic features (GeoNames, 2018) that match the description. If multiple matches are found, the midpoint and the minimum distance between the matches are calculated. There were many accessions for which this approach only matched the term “Java” and hence returned the midpoint of the island. As this would heavily bias the spatial patterns those were excluded from further analysis.

As recommended by Hernandez *et al.*⁸¹ the species specific distribution models were only built for species for which we had at least five unique occurrences, which was true for 1,129 species. Modelling the species distribution which such a low number of samples is only possible based on the limited area for which they are predicted (island of Java)⁸². Since multiple species had to be modelled, the stacked species distribution model implemented in the R package SSDM⁸³ was used. For each species an ensemble species distribution model is created, meaning that the distribution is calculated by different algorithms and a consensus in the resulting species distribution models is found based on an average of the models. For this all commonly used algorithms (general additive models, generalized linear models, multivariate adaptive regression splines, classification tree analysis, generalized boosted models, maximum entropy, artificial neural networks, random forest, support vector machines) were used. Since also absence data are necessary for the modelling and such are not available for the plants, pseudo-absences were selected according to Barbet-Massin *et al.*⁸⁴ also implemented in the SSDM package. For the regression algorithms (generalized linear model, generalized additive model and multivariate adaptive regression splines) and maximum entropy 1,000 and 10,000 pseudo-absences are chosen randomly independent from the number

of presence data. For the remaining algorithms, the number of random absence data is equal to the number of presence data, unless the number of presence data is below 100, than the square of the presence data is used. Both presence and absence data are weighted equally for all algorithms in the model prediction. The modeling of each algorithm was repeated 20 times and cross-validated by the “k-folds” method. Here the presence and absence data for each species is partitioned into k-1 training sets and one evaluation set. For the modeling of the Javanese plants k=4 was chosen. The region for which the species distribution is modeled is divided into a raster of several grid cells (10 * 10 km). For each grid cell, a probability is calculated if the species occurs there and is converted into presence/absence data by applying a threshold that maximizes the true skill statistic (TSS)⁸⁵. The presence data is afterwards stacked to obtain the species richness data. The modeling is based on climate and soil data, also on a 5 arc-minute resolution (approx. 10 km) as environmental predictors. The Worldclim project⁸⁶ provides bioclimatic variables, while SoilGrids database⁸⁷ makes soil data accessible. The respective data for Java were downloaded from the sources and rescaled to 5 arc-minute resolution (a list of all variables can be found in Table A5). For the evaluation of the stacked species distribution model, six evaluation metrics are calculated as described by Pottier *et al.*⁸⁸ and implemented in the SSDM package. The species richness errors describe the difference between the predicted and observed species richness. While the assemblage prediction success is the proportion of correctly predicted species distributions, the assemblage kappa gives the proportion of specific agreement. The amount of true negatives, that are species that really do not occur in the respective gridcell, is given by the assemblage specificity, whereas the assemblage sensitivity gives the true positives, meaning the species that are observed and predicted for the respective grid cell. The Jaccard index is a metric describing the similarity and diversity of communities.

Having the stacked species distribution models in hand, one can visualize the species richness as a representative for the biodiversity. However, the initial aim was to study the chemodiversity of Java. Therefore, a Shannon index was calculated for each grid cell by Equation 2.

$$H = \sum_{i=1}^S p_i \cdot \ln p_i \quad (2)$$

With p_i being the proportion of a metabolite i to the total number of metabolites in a

grid cell and S being the individual metabolites present in a grid cell. As metabolites, all secondary metabolites for the plant species from Java were used, collected as described in the passage “taxonomic and metabolite information”.

2.2.5 Habitats

To address the question in which habitats or ecoregions plants with most identified natural products grow, all available occurrences on 1 April, 2019 were downloaded from GBIF⁷¹. The following issues and flags were accepted, as they can be assumed to have no influence on the planned analysis:

- Coordinate rounded
- Geodetic datum assumed
- WGS84 Coordinate reprojected
- Country derived from coordinates
- Taxon match higher rank
- Multimedia date invalid
- Multimedia uri invalid

The issues regarding the coordinates or country were accepted, because this analysis is based world wide and further abstracted in the following steps. Therefore, an as exact occurrence point as with the analysis based only on one island, it was not necessary. Multimedia flags were included, since the multimedia data is not used in this study. The coordinates were used together with the WWF terrestrial ecoregions⁷² to assign habitats and ecoregions to the plants. Afterwards it was evaluated in which habitats most of the plants were growing, for which of them metabolite information is available, and what is the total number of different metabolites per habitat.

2.3 Results and Discussion

2.3.1 Database crawling

With the database approach, we had in total 29,981 plant species accepted by The Plant List, 1,288 unresolved species and including synonyms 38,794 species in total which are either from the Java Plant list contributed by our cooperation partners or referenced as Indonesian plants by GBIF or KNApSAcK.

With this list of species the KNApSAcK database was crawled for information about

natural products in those plants. The result is summarized in Table 2.1. Since this is a quite small amount of species covered by this search and a comparison to the Reaxys database revealed that a lot of metabolites are missing, we also searched in the Dictionary of Natural Products (DNP). As shown in Table 2.1 in the DNP more species are covered and more metabolites are referenced compared to the KNApSAcK database. But still there is a lot of information missing, as the comparison to a manual literature search for 20 species reveals. However, one important point of the literature search with 20 plant species was, that although KNApSAcK and DNP do not contain all natural products, which can be found in literature, it seems that all plants which have no metabolite entry in KNApSAcK or DNP, the literature search also does not find publications with natural products of those plants.

Table 2.1: Summary of results gained by crawling the KNApSAcK database and the Dictionary of Natural Products (DNP).

	Species with referenced metabolites	Species with bioactive metabolites	Metabolites in total	Bioactive metabolites
KNApSAcK	2,065	1,105	15,796	2,101
DNP	3,953	2,066	45,607	6,280

2.3.2 Chemical Diversity

The statistics of the species growing in the different habitats (see Figure 2.1) show that most of the species are growing in the moist broadleaf forest, since this also the most widespread habitat in Indonesia. However, the highest percentages of (active) metabolite containing species can be found more in the dryer habitats like grasslands, savannas and shrublands. Since there are regions in Indonesia where more explorations have been done for identifying plants, it is reflected in the pattern of plant occurrences downloaded from GBIF⁷⁹ (see Figure 2.2A). There are some hotspots where more plant records are stored than in other regions.

Although the amount of data has to be increased, the existing information based on KNApSAcK and GBIF has been used for first analyses. Looking at the GPS referenced plant species it is obvious, that there are hotspots of plant occurrences. This is due to the fact that most research projects are locally limited. However, since most of

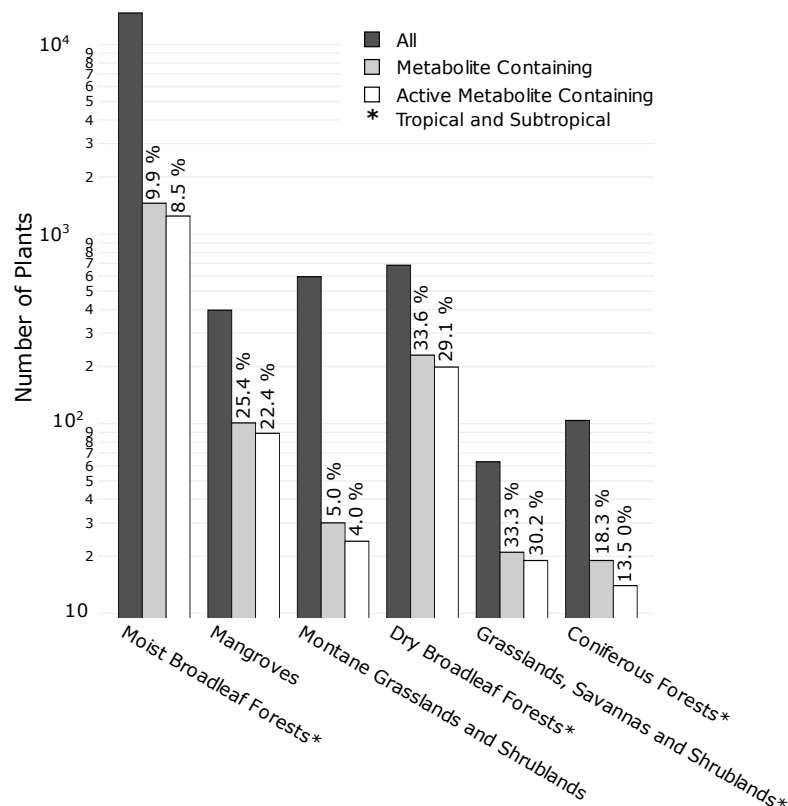


Figure 2.1: Half logarithmic representation of the number of Indonesian species according to GBIF coordinates per WWF habitat. The different colors indicate the total number of species, the number of metabolite and active metabolite containing plant species. Together with the percentages of plants in each category for which metabolites are recorded in the databases.

Indonesia is tropical and subtropical moist broadleaf forests, most species with GBIF GPS references are located in this habitat (17,310). This habitat is followed by dry broadleaf forests (1,063), grasslands (947), mangroves (550), savannas (127), coniferous forests (99), mixed forests (41) and desert (14). Although for the desert the lowest number of plants is found, most of them contain metabolites (14 %, 2-4 % in the remaining habitats) as well bioactive metabolites (7 % vs. 1-3 %, resp.). But looking on those percentages, we have to keep in mind that the information about metabolites is still limited, since this is only based on data from the KNApSAcK database. Based on the existing data it seems, that in mangroves those plants containing natural products (based on KNApSAcK), have the highest amount of natural products per plant (18.5

total metabolites per plant, 1.3 active metabolites per plant), whereas the plants from coniferous forests and desert seem to contain a higher amount of active metabolites (1.5 active metabolites per plant each, 13.3 and 7.5 total metabolites per plant respectively). But these analyses have to be looked upon with caution, since it is just an excerpt of information, which is available.

The major part of the reported species occurrences on Java are located around the big cities, roads and the protected areas (Figure 2.2A), obviously due to the easier access to these regions or sample campaigns. Those campaigns mainly focus on protected areas, because the major part of Java is used for agriculture. The protected areas are the small regions where still a wide biodiversity can be found. The distribution of 1,132 plant species was modelled on the island of Java. The soil data had a higher influence on the modelling than the bioclimatic data (see Appendix). The highest species richness (Figure 2.2B) can be found near the mountains of Java. As Figure 2.2E shows, the number of metabolites in a grid cell is correlated to the number of species. Therefore, it is not surprising that the chemical diversity (Figure 2.2C) corresponds to the species diversity (Figure 2.2B). One would assume, that such a correlation curve is not linear, but that it flattens at a certain number of species like a saturation curve. However, since we only look at the island of Java with a rather small number of species compared to the world with many similar habitats (and likely convergent chemical evolution), we are most probably in the first region and therefore still in the linear rise part of such a saturation curve. The ratio of the chemical diversity (Shannon Index) per the number of plants in a grid cell shows the regions of Java were plants with a higher chemical diversity grow (Figure 2.2D), in contrast to the overall chemical diversity (Figure 2.2C). Here not the mountainous regions, but the flatter regions seem to be the more interesting ones (south of the island between Surabaya and Semarang, and the northern coast line). However, also here great care must be taken, since plants of this region are usually better studied than less accessible, less abundant mountain species.

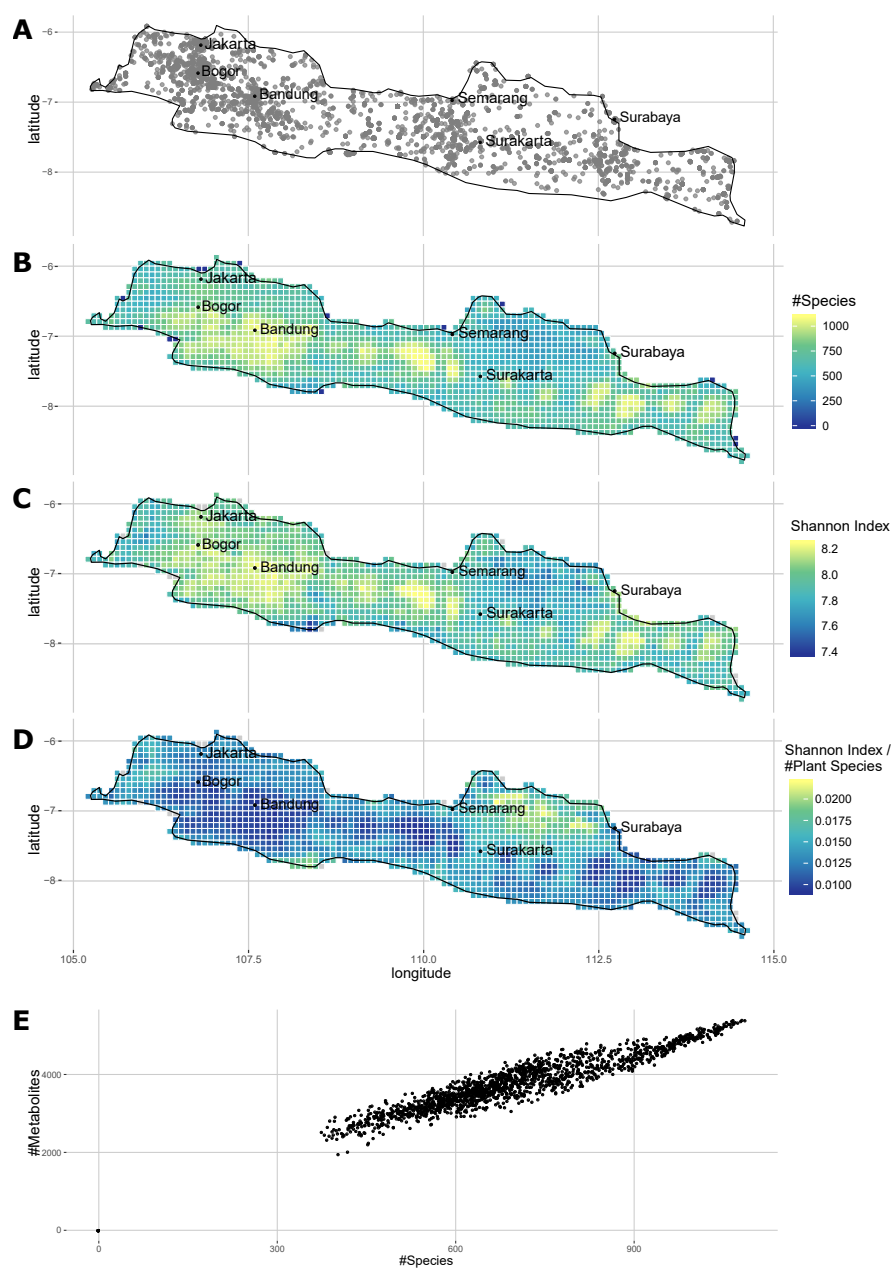


Figure 2.2: **A-D** map of the island of Java with the six biggest cities. **A** GBIF occurrences of plants species which have at least four occurrences on Java. **B** modelled species distribution. **C** the chemical diversity represented by the Shannon index. **D** ratio of the Shannon index per the number of plant species in each grid cell. **E** the number of metabolites versus the number of species per grid cell of Java.

2.3.3 Metabolite information

We used metabolite information from the KNApSAcK and NPASS databases for 1,996 plant species growing on the island of Java to develop a workflow for identifying natural sources and products worth to be studied in detail. The 10,612 and 9,894 metabolites were compared for identification of duplicate structures or names and were classified as primary and secondary metabolites via pathways of the KEGG database. In total we have 16,503 unique secondary metabolites deriving from the two databases. In the KNApSAcK and NPASS database also bioactivities are accessible for the metabolites. Those were also stored and categorized into 24 groups following the economic botany collection standard (EBDCS). The main interest was antiinfective activities (including anti-fungal, anti-microbial, anti-parasitic and anti-viral). Those were further reduced to only contain those activities affecting vertebrate pathogens resulting in activity information for 1,640 species.

In figure Figure 2.3 the natural product classes of the Javanese KNApSAcK/NPASS dataset are shown together with the total number of antiinfective metabolites in each class and colored by their amount of all metabolites in the respective class. One clearly sees, that although quinones (289 metabolites) are relatively uncommon metabolites, ca. 30 % of them have antiinfective properties (85), which is only higher in the very rare fluorenes, where in total only eight metabolites are reported. In contrast, the class of sugars contain the lowest proportion of antiinfective metabolites (in relation to the class size).

2.3.4 Phylogenetic Diversity and Taxonomic Analysis

The analysis of the phylogenetic patterns identified 17 clades across the entire phylogeny with a significant degree of clustering (i.e. overrepresentation of antiinfective activities, 'hot nodes'). The family of Meliaceae is for example contained in such a cluster. The taxonomic analyses (Figure 2.4 and Table A4) showed that the average ratio of antiinfective species per family was 26 % and 20 % for the analyses of all taxa and native taxa only, respectively. When compared with the phylogenetic analyses, we found a large degree of overlap. Like in the phylogenetic analyses, Amaryllidaceae and Asparagaceae lay above the 95 % CI (confidence interval) when all species were considered but fell within the CIs when only the native species were analyzed. The fourth family in this clade, Liliaceae, however, did not fall above the 95 % CI in both

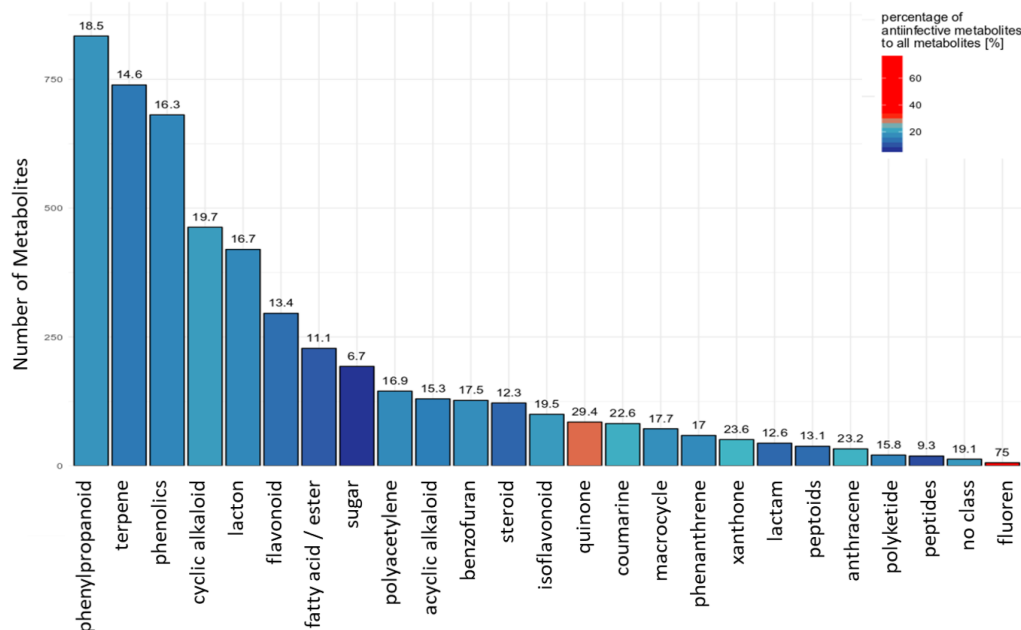
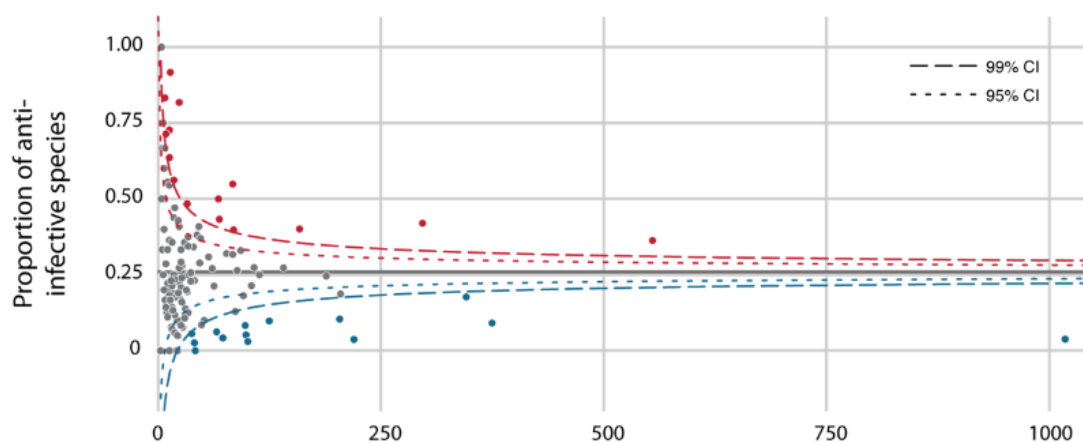


Figure 2.3: Natural product structural classes identified in the Javanese KNApSACk/NPASS dataset. The height of the bars indicates the number of metabolites with antiinfective activity in the respective class. The percentages on top of the bars and the colors indicate the relative amount (percentage) of antiinfective metabolites to the total amount of natural products (antiinfective, other activity, and inactive) in this class. The class “phenylpropanoid” has the highest number of antiinfective metabolites in the data set, while the class “quinone” has a quite low total number of metabolites, but 29.4 % of all quinones in the data set are antiinfective.

analyses. We also found a few notable differences between the analyses, as a number of families that are above the 95 % CI in the taxonomic analyses do not appear at all (e.g. Polygonaceae and Rhizophoraceae) or only partially (Fabaceae) in the “hot clades”. In some cases, the phylogenetic analyses revealed a more complex pattern of over- and underrepresentation of antiinfective activities than the taxonomic analyses. Rubiaceae fell below the 95 % CI in the taxonomic analyses (despite having a number of species with documented antiinfective activities), while in the phylogenetic analyses two small “hot clades” are recovered with 11 and 9 species, respectively. These all belong to the subfamily Cinchonoidae (tribe Naucleae), which is characterized by a significant occurrence of oxidized indole alkaloids. Thus, the low percentage of species described as antiinfective in the taxonomic analysis is due to the restriction of antiinfective activities studied or found in only one subfamily within the very species-rich Rubiaceae.

(A) All plant species



(B) Native plant species only

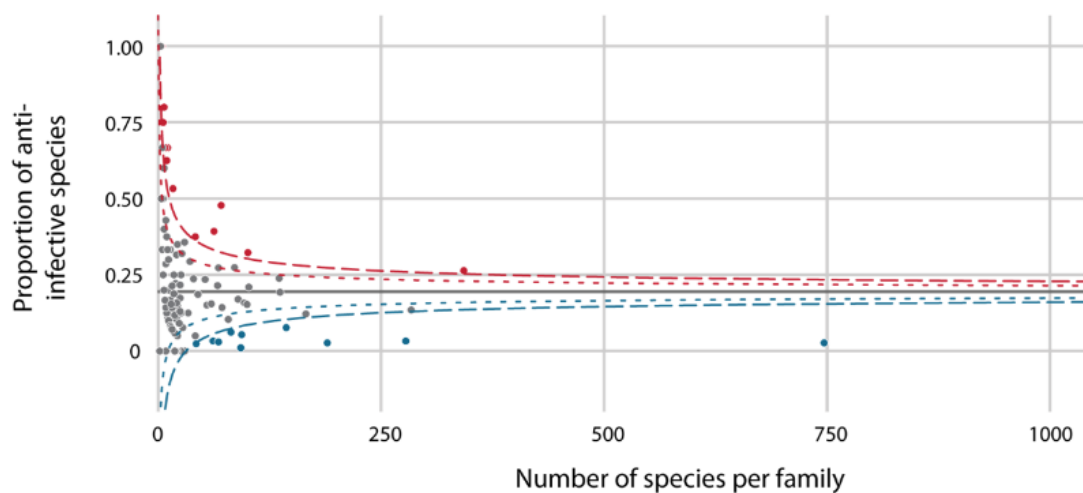


Figure 2.4: Taxonomic analysis of antiinfective effects in seed plant species across the flora of Java. For each plant family, we calculated the proportion of species with antiinfective effects (metabolites with known activity). The overall mean of all families was then used to calculate the 95 % and 99 % confidence intervals (CI) depending on the number of species per family for (A) all seed plant species recorded on Java, and (B) native seed plant species only, respectively.⁵⁴

2.4 Summary

With the presented approach, strong phylogenetic and spatial patterns were detected in the distribution of antiinfective activities across the flora of Java. This proof of concept indicates the combination of phylogenetic, spatial and phytochemical information can be a usefool tool to guide the selection of taxa for directing efforts of natural product extraction. Species from “hot clades” without documented natural products or bioactivities have a high probabilitiy of chemical compounds and activities. Importantly, given the sometimes considerable variation of chemical compounds among closely related species⁸⁹, these species might still provide new compounds or activities. While species from “cold clades” have an overall low probabilitiy for the examined bioactivity (here “antiinfective”), but at the same time any compound found has a higher probability of being new for the screened activity, and potentially even structurally novel. Therefore, it is of high interest to apply this approach globally. In the following it is described how a global dataset was created and analyzed.

3 Creation of a Species - Natural Product Gold Set

3.1	Material and Methods	25
3.1.1	Publication collection	25
3.1.2	Preprocessing of the corpus	26
3.1.3	Annotation of the corpus	27
3.1.4	Entity annotation guidelines	27
3.1.5	Relation annotation guidelines	29
3.1.6	Chain reasoning	32
3.2	Results and Discussion	32
3.3	Summary	36

A gold standard set is a collection of literature in which all relevant named entities and relations are annotated. Therefore, a gold set is always issue specific and can hardly be used for other purposes. The gold standard corpus used for the following development of an information extraction tool was created together with the OntoChem GmbH (formerly OntoChem IT Solutions). In the gold set named entities and relations already defined by OntoChem were used. Additionally the relation “species produces natural product” as a semantic concept comprising everything from which one can conclude that a natural product is produced by a species for example by trigger words like “isolated from” or “metabolized to” was defined. Natural products in this context are secondary metabolites like small molecules or oligomers/polymers such as polysaccharides, but excluding proteins and genes. Primary metabolites maybe included if they are specifically mentioned in the base set of the papers. However, since primary metabolites are present in all organisms, and only relevant for, e.g. metabolomics if quantified, they are not in the focus of our information retrieval. Concerning the

species, we focused especially on Indonesian plants which is reflected by selecting appropriate publications. Nevertheless, we consider the papers chosen as typical for natural product papers, independent of any geography or organism kingdom. Thus, also for species from the other kingdoms (Monera, Protista, Fungi and Animalia) the “species produces natural product” relation was annotated if it occurs in the texts. In the gold set such relations are called `species_Produces_compound` or more specifically, e.g., `plant_Produces_compound`. In the following the publication collection and all annotation rules are explained in detail.

3.1 Material and Methods

3.1.1 Publication collection

For creating the gold standard corpus the first step was the selection of relevant publications. Therefore, the OC|miner developed by the OntoChem GmbH was used, which is a text analysis and data mining toolbox. Due to underlying ontologies for the search function, it is possible to include roles in a semantic search⁹⁰. Entering a term like “plants” as search concept from the “species” ontology, the program does not only search for the term plants, but also for all children of “plants” in the ontology “species”, which are plant families, genera and species. As we are not only interested in the plants and their natural products, but also in bioactivities, and methods how the natural products have been isolated, terms correlating to these topics were also included in the query. Since our special interest is on Indonesian plants, the species search term was not only “plants” but “Indonesian plants”. Indonesia is one of the world’s top three diversity hotspots, representing a wealth of different genera and habitats, producing a large array of different natural products, requiring different isolation techniques. “Indonesian plants” is a node in the species ontology to which plant names based on the Flora of Java⁵⁵ are linked. The actual query was:

```
+spec:"Indonesian plants" +(eff:inhibition eff:antimicrobial eff:anti-inflammatory dis:) +(t:"phytochemistry" t:"phytochemical" t:"pharmacology" t:"pharmacological" t:"natural product" t:"natural products") +chem: +meth:assay +meth:extraction -meth:"in silico methods"
```

This query is based on the following rules:

- “+” word of this annotation type has to be present (specification or *= all of this

type)

- “-” none word of this annotation type has to be present
- (nr1 nr2) nr1 and/or nr2 has to be present in text
- spec=species; eff=effect; t=free text; chem=chemistry; dis=disease; meth=method

This query resulted in 8,300 documents in the PubMed Central open access database of full text publications⁹¹. These documents were ordered by relevance and the top 100 list items were selected for the creation of the gold corpus.

To have a more balanced corpus for testing false positives, we also added publications which are not correlated to the topic of natural products. Therefore, eight publications and one part of a patent of existing gold corpora at OntoChem IT solutions were added. Two more additional publications were added that have been found with the following OC|miner queries:

```
t:"heavy metals" +meth:"analysis" +spec:"Indonesian plants" +t:"accumulation"  
t:"pesticide" +meth:"extraction" +spec:"Indonesian plants" +t:"accumulation"
```

With these queries, two publications were selected: one dealing with pesticides in plants and one with heavy metals in plants. Thus, in total this gold standard corpus consists of 102 documents with 100 documents focusing on the “species produces natural product” relations and two additional texts dealing with pesticides and heavy metals in plants to gain a balanced corpus by using publications dealing with different topics in the same corpus.

3.1.2 Preprocessing of the corpus

As a starting point for the manual annotation of the corpus, the selected 102 documents were annotated using the standard automated annotation pipeline of OC|miner⁹⁰. This high-performance text processing system is based on Apache UIMA and uses a range of different annotators that are created from domain specific ontologies. Therefore, the resulting starting text annotations also contain named entities from other domains such as plant structures, effects, methods, or clinical trials. Since these are of minor importance for the species – natural product relations, they were ignored in the subsequent processing of the text if it was found that they are correct. If those entities were incorrect, they were labeled as such.

The corpus was split into two subcorpora, the training* and test corpus, consisting of 50 and 52 publications, respectively.

3.1.3 Annotation of the corpus

Since the task of extracting “species produces natural product” relations from texts depends on the correct annotation of species and chemistry, we decided not only to annotate those species and chemistry terms that are involved in a relation, but all species and chemistry terms occurring in the publications. This enables the development or improvement of named entity recognition systems.

After the preprocessing of the 102 selected full text publications, they were annotated manually according to annotation guidelines using an annotation tool called brat⁹². The initial guidelines were constructed to extract “species produces natural product” relations from text and are based on the annotation pipeline of OC|miner. However, due to unspecific or ambiguous text phrases the guidelines had to be adjusted and therefore the improvement of the annotation guidelines became an iterative process in close relation to the annotation process. The final annotation guidelines are listed in a summarized way below.

A domain specialist performed the annotation in close agreement with an expert for creating gold standard corpora and for the development of text mining systems. Therefore, we cannot calculate an inter annotator agreement.

3.1.4 Entity annotation guidelines

In accordance with the aim of the corpus, the annotation guidelines were defined to also cover “species produces natural product”-relations spanning beyond the borders of sentences, paragraphs or tables. Therefore, all species and chemistry entities had to be annotated as such as well as their relations.

General entity annotation guidelines:

- Generally only nouns or pronouns are annotated, adjectives shall not be annotated
- Only chemistry and species entities are examined
- Anaphoric terms like pronouns relating to chemistry or species entities have to be annotated as anaphor and have to be connected to the respective entity via the relation Coreference

- Labels, acronyms and abbreviations of species or chemistry entities have to be annotated as the corresponding entity and marked as label, acronym or abbreviation. The long form has to be added as “AnnotatorNote”
- Pre-annotations have to be checked if they are correct and if necessary have to be corrected

Species annotation guidelines:

- All species are annotated with the entity “species”
- The species annotation is based on OntoChem’s species ontology
- According to their affiliation, species entities are assigned as plant, fungus, alga or bacterium in the corresponding species ontology
- If it is not possible to allocate the entity to one of those groups, it is labeled as species
- Latin, trivial or locally used names and abbreviations of species, genera and families are labeled
- Always the longest form has to be annotated (*Aloe vulgaris* Lam., not *A. vulgaris*)
- In enumerations of species, each species entity has to be annotated on its own
- Specifications also have to be annotated on their own e.g. *Curcuma longa* (Zingiberaceae), the species as well as the family have to be treated as separate entities
- Another entity used is HerbDrug, describing formulae especially used in traditional medicine containing plants

Chemistry annotation guidelines:

- Chemistry entities are divided into chemical Compounds, Classes or Groups (chemComp, chemClass and chemGroup), Substances and Polymers
- The chemistry annotation is based on ontologies and databases of OntoChem
- chemCompound is used as entity for specific chemical compounds, which are chemical names to which a certain defined and – more or less unique * structure can be matched (e.g. curcumin)
- chemClass refers to chemical classes, which are chemical terms describing families of chemical compounds and it is not possible to assign a certain structure (e.g. steroids)
- chemGroup are chemical entities describing small parts (moieties, substructures, substituents etc.) of a bigger structure (e.g. methyl group)

- Substances are general or functional concepts like “constituents”, “extract”, “inhibitor” or “mixtures”
- Polymer refers to chemical compounds consisting of several repetitions of smaller units (e.g. algin), excluding peptides (>16 amino acids)/proteins and RNA/DNA
- Peptides consisting of up to 15 amino acids are annotated as chemCompound
- Enumerations of chemistry mentions are tagged as separate entities (e.g. sesquiterpenes Dictyophorine A and B are annotated as one chemClass sesquiterpenes and two chemCompounds Dictyophorine A and Dictyophorine B)
- Enumerations of labels (e.g. 2-5, where 2, 3, 4 and 5 are labels for compounds) are annotated as chemCompound and as note “Container: “ followed by the enumeration of the compound names is added
- The term “compound” is annotated as chemClass although it is an unspecific term, since it normally refers to a chemCompound mentioned actually elsewhere in the text

Natural Product flag guidelines:

- Chemical mentions are only flagged as NaturalProduct if the context marks them as natural
- Contexts marking a chemical entity as natural products are for example “compound was isolated from a species”, “the natural product compound”, “compound biosynthesis” or “compound is produced by species”
- Contexts not specifying a chemical entity as natural products are for example “compound exhibits antibacterial activities” or “compound is metabolized”
- Natural products can only be small molecules or polymers such as polysaccharides and must not be proteins or genes, since these are from the view of natural product chemistry separate domains

3.1.5 Relation annotation guidelines

The aim is to annotate relationships between species and chemicals which indicate that the chemical was produced by the species and is therefore a metabolite or natural product of the species, e.g. “Withaferin A was extracted from *Withania somnifera* leaves” or “Withaferin A can be isolated from *Withania somnifera*”, as plant_Produces_compound relation. Relationships such as “Withaferin A can be produced in a transgenic *Arabidopsis* line” or “Gibberellin treatment of *Arabidopsis thaliana* tissue” should not be

annotated. For relations such as “the pesticide Picoxystrobin was found in species” the relation `species_Contains_compound` was used.

In principle, no other relations were used. However, since we also want to annotate relations beyond one sentence and within tables we also used the relations `isInstanceOf` and coreference to gain a close proximity to the text. Another reason for using this fine-grained annotation are the long distances between partners of the “species produces natural product” relations, which are often relation chains rather than direct relations.

Especially in relations that span beyond one sentence one partner is often represented by a pronoun e.g. “XY is an important medical plant. Compound ZZ was isolated from it.” These should be connected to the noun they represent by a coreference relation. If there is an enumeration of plants or chemistry, those can also bear relation to the other, e.g. “plant species (plant family)” here the species is part of the family or “chemistry class, such as compound 1 and compound 2” where compound 1 and 2 are belonging to the chemistry class. In such cases the children should be connected to the parent via an `IsInstanceOf` relation. The exact as well as some general guidelines for the different relations are listed below.

General relation annotation guidelines:

- Relations can be located within one sentence or table or can span several sentences
- The direction of the relations (starting and endpoint) are crucial * the respective directions are given at the description of each relation
- `IsInstanceOf` and coreference relations are only used if one of the partners participates in a “species produces natural product” relation
- `IsInstanceOf` is used to connect children of a domain to their parent (see Figure 1). The relation always starts at the child and ends at the parent
- Coreference relations connect entities describing the same, e.g. an anaphoric term and its antecedent (see Figure 1A). Especially pronouns and their respective nouns are connected. This relation always starts at the anaphoric term and ends at the antecedent.

Species - Chemical relation annotation guidelines: There are several relations that can be used, dependent on which species entity is part of the “species produces natural product” relation:

- `plant_Produces_compound`

- fungus_Produces_compound
- alga_Produces_compound
- bacterium_Produces_compound
- species_Produces_compound

All of these relations are annotated in accordance to the same following guidelines

- The relation always starts at the species and ends at the chemical entity (see Figure 1)
- If one of the partners is also part of an isInstanceOf relation, it has to be the endpoint of this (see Figure 1A)
- If one of the partners is also part of a Co-reference relation, it has to be the starting point of this (see Figure 1A)
- Only chemical and species entities are connected which are clearly in a “species produces natural product” relation, therefore especially the context is important, e.g. “the pesticide Carbendazim was isolated from plant species” does not indicate a plant_Produces_compound relation
- General terms of chemical or species entities such as compound, constituent, pigments, plant, bacteria or endophytes are allowed in the relations
- Proteins and genes must not be contained in the relations, since these are different domains from the view of natural product chemistry
- For relations that contain the entity HerbDrug the relation plant_Produces_compound was used

To demarcate the “species produces natural product” relation from a context, which indicates that a chemical can be found in a species but is not produced by this one the following relations were used:

- plant_Contains_compound
- fungus_Contains_compound
- alga_Contains_compound
- bacterium_Contains_compound
- species_Contains_compound

Sentences like “the pesticide Carbendazim was isolated from plant species” are annotated with this type of relations to be distinguished from the natural product relations. The annotation guidelines for these relations are similar to the ones described above:

- The relation starts at the species and ends at the chemical entity
- Here the context has to suggest that the chemical entity can be found within the species, without being produced by it

3.1.6 Chain reasoning

To gain a gold standard corpus only consisting of the relations we are interested in (species – natural products/metabolites), a chain reasoning is performed according to the gold standard corpus of Schlaf *et al.*⁹³. Herein, all `isInstanceOf` and coreference relations are replaced by “species produces natural product” relations with the respective partners as shown in Figure 1.

3.2 Results and Discussion

The final gold standard corpus is based on 100 full text publications dealing with the extraction of natural products from different species and two publications about extraction of pesticides and heavy metals from plants, respectively. The publications were selected as described above from the PubMed Central database⁹¹. Due to the query which was used to create the corpus, the publications mainly deal with natural products isolated from plants with biological activities, such as antimicrobial properties. The corpus was split into a training set consisting of 50 publications, and a test set with 52 publications. In total, the corpus contains 106,250 sentences, 50,995 in the training set and 55,255 in the test set.

The corpus includes in total 39,729 chemical and 16,166 species entity mentions. The biggest group of chemical entities, the `chemCompound` (49.6 %), is also of major importance, since these are constituents (metabolites, chemicals) with a defined chemical structure which are of specific interest in “species produces natural product” relations. The chemical entity types `chemClass` and `Substance` contribute 23.8 % and 20.6 %, respectively, to the total number of chemical entities while a minor contribution comes from `chemGroup` (2.6 %) and `Polymer` (3.3 %). Beside `chemGroup`, all chemical entities can be part of “species produces natural product” relations. However, only 21.8 % of all chemical entities are additionally labelled as natural products, since not all contribute or are part of a “species produces natural product” relation. For example, found in a specific context, but can also be mentioned as solvents used, or as reference compounds. The exact distribution of “species” and “chemistry” entities in the single sets and the

Table 3.1: Number of entity annotations in the training- and test set as well as the deduplicated counts per set and in Total.

	Training set	Training set deduplicated	Test set	Test set deduplicated	Total
Species	9,254	2,600	6,912	1,812	16,166
chemClass	5,140	2,052	4,309	1,675	9,449
chemCompound	8,871	3,837	10,842	3,375	19,713
chemGroup	506	320	540	371	1,046
Polymer	470	196	854	231	1,324
Substance	4,489	1,218	3,708	1,192	8,197
Total	79,476	7,623	20,253	6,844	39,729
“Chemistry”					
Natural Products	5,897	2,748	4,195	1,352	8,645

total corpus can be found in Table 3.1. Here also the numbers of the deduplicated entities per set are listed, i.e. each unique relation is only counted once. On average, this is true for one third of the total number of entities for chemicals. This shows the diversity of the chemicals mentioned in the publications used for the creation of the gold standard corpus. For the entity “species”, the deduplicated amount is about 26 % of all species per set. This is because most publications focus on one species, genus or family and a single publication is therefore more limited in the amount of different species, while each of those can produce diverse natural products. The training set contains more deduplicated total chemistry entities (sum of chemClass, chemCompound, chemGroup, Polymer and Substance) as well as deduplicated species entities.

After the chain reasoning, the corpus comprises a total of 8,079 “species produces natural product” relations. As expected and presumed at the beginning, most relations (54.9 %) cannot be found within one sentence, but are spanning over multiple sentences or are located within tables (see Table 2). The training set contains about twice as many “species produces natural product” relations than the test set. Most of the “species produces natural product” relations are covered by plant_Produces_compound relations (86.0 %), since the publications on which the corpus is based were selected to predominantly contain these relations. The number of deduplicated relations are about

half of their total number and show that although there are some redundancies, the publications contain a quite diverse set of relations.

We created a manually annotated gold standard corpus, whereby the corpus is gold in terms of chemical entities, species entities and “species produces natural product” relations. It is based on 102 recent publications being typical for the subject area. The set consists of 106,250 sentences, 16,166 species entities, 39,729 chemical entities and 8,079 “species produces natural product” relations. It is therefore comparable to the ChEBI corpus⁴⁹ in size and number of species and metabolite entities, but additionally contain the feature relations. However, to our knowledge it is the first gold standard corpus annotating relations spanning over multiple sentences and connecting to tables. Table 3.2 clearly shows that about the half of all relations cannot be found within one sentence, but are spanning over multiple sentences. Therefore, our gold standard corpus closes the gap of fragmentary relation annotations and offers great opportunities to train new information extraction tools dealing with more complex relations covering the whole available information from one publication.

Dealing with the named entity recognition of chemical entities, there are some difficulties which have to be overcome^{39,42}. First of all, there can be many names describing the same chemical structure by using trivial, IUPAC or trade names. If these names as well as their synonyms are stored within a chemistry database, the assignment can be performed unequivocally as a semantic concept. Such chemical ontology concept is typically based on a given chemical structure that is assigned to the chemical named entity. Accordingly, the ID of the ontological concept (OCID - ontology concept id) is assigned to the chemical entities, and the same OCID is also added to the synonyms present in the text or even in different publications. If a chemical entity is not present in the database, it is annotated manually. In this case, synonyms used in the text can be connected via the relation “synonym of”. However, as soon as several texts dealing with the same structure use different names not stored in the database, they are not treated as one chemical concept. This has to be kept in mind when using the corpus as a basis for the development of text mining tools.

Furthermore, one expression can refer to different things, e.g. “antioxidant” can be annotated either as an effect or as a chemical (compound) class. Here the annotator had to decide to which named entity the context pretends.

Since next to sentences indicating a “species produces natural product” relation such

Table 3.2: Distribution of annotated “species produces natural product” relations per set, as well as the total number and the distribution within the text. “Spanning over sentences” contains relations spanning more than one sentence, as well as relations located in tables.

	Training set	Training set deduplicated	Test set	Test set deduplicated	Total
Algae	576	439	222	124	798
Bacteria	19	19	118	71	137
Fungus	27	20	111	86	138
Microbe	3	1	22	17	25
Plant	4,461	2,744	2,486	1,247	6,947
Other	1	1	33	29	34
Species					
Total	5,087	3,224	2,992	1,574	8,079
Relations					
Within one sentence	2,172	-	1,473	-	3,645
Spanning sentences	2,915	-	1,519	-	4,434

as “Withaferin A was isolated from *Withania somnifera*” there are also semantically similar sentences like “the pesticide Carbendazim was isolated from plant species” which clearly do not indicate a “species produces natural product” relation due to the word “pesticide”. Those relations are annotated as “species contains chemistry”. Having those relations also annotated it is easy to check the quality of the information extraction tool annotations if those also contain such relations as false positives for “species produces natural product” relationships.

To gain a broader application field for our gold standard corpus, we did annotate all mentioned species and chemistry entities. Consequently, the corpus can be used for the development or improvement of NER (named entity recognition) tools, either rule based or machine learning tools.

Due to ongoing efforts made on natural product extraction and identification, one can expect the continuous occurrence and creation of new trivial names for these secondary

metabolites. Therefore, pure dictionary or ontology based NER hardly can annotate those new natural products correctly. Consequently, rule based or machine learning (ML) approaches are expected to promise to be more efficient for these cases since they consider the context or syntax of a chemical entity. To make our gold standard corpus also useful for the development of such ML tools, each chemical entity within a context describing it as a natural product (see annotation guidelines) was labeled with the flag “naturalProduct”. The same chemical entity outside of a natural product context was not labeled as natural product, in order to not provide misleading context untypical for natural products to the ML algorithm. For example, in a sentence like “Withaferin A has anti-inflammatory properties” the compound Withaferin A could be a natural product but could also be some synthetic compound. Therefore, in such mentions of presumable but not safely identifiable natural products, they were not labeled as such.

The fine grained relations, as well as the relations spanning over sentences and including tables make the IPB gold standard corpus unique and shall allow to develop novel information extraction tools. These tools can be used to create and update databases containing species and their respective natural products. This in turn will be a great resource for all sciences involving studies of Nature's chemistry like, e.g., natural product based drug development, metabolomics, chemical ecology, or food and herbal drug quality research.

The final gold standard corpus is based on 100 full text publications dealing with the extraction of natural products from different species and two publications about extraction of pesticides and heavy metals from plants, respectively. Due to the query which was used to create the corpus the publications mainly deal with natural products isolated from plants with biological activity, such as antimicrobials. The corpus was split into a Training Set consisting of 50 publications, and 52 publications in a Test Set. In total these are 106,250 sentences, 50,995 in the Trainings Set and 55,255 in the Test Set.

3.3 Summary

With the development of a gold set for plant produces natural product relations it was possible to program a suitable tool for the relation extraction from publications. Although the tool has still room for improvement a quite high number of relations, which is comparable to existing databases, was extracted from PubMed Central open

access articles. Now the relations have to be evaluated and further analyses on this higher amount of data has to be done.

Here we present the creation of a gold standard corpus, the “IPB organism-metabolite corpus I”, for species and chemical entities as well as for “species produces natural product” relationships. Furthermore, chemical entities which are natural products are additionally labeled as those. The corpus consists of 102 full text publications available at PubMed Central. It is the first gold standard corpus where relations are not only annotated within sentences, but also span beyond sentences and within tables. In the future we plan to extend the corpus by expanding the organism base which is currently plant biased, but especially by adding additional entities, such as effects, diseases, proteins or plant structures (organs). These will also be annotated manually to get a gold standard for those entities too. Additional relations such as “compound treats disease” or “compound has effect” could be added. Different languages other than English are also of importance in the Natural Product field, as much information on natural products is found in other languages from traditional Chinese medicine, via Sanskrit, Arab or European (German, Spanish, French etc.) papers. The corpus can be used to develop novel text mining tools. Its main purpose is to train natural product entity recognition systems or an information extraction tool for “species produces natural product” relations. Especially for natural products there is no satisfying named entity recognition system, since the names of natural products are very diverse, there is no central register, and regularly there are new natural products found and named. Therefore, neither a dictionary lookup system nor a chemical entity tagger developed for patents or pure chemistry publications can be used with success. Therefore, this corpus provides the currently best basis for developing and evaluating new text mining tools in the field of natural product research. We also hope it can contribute to a better comparison of tools developed, to show the strengths and weaknesses of the different approaches.

4 Creation of a Global Data Set

4.1	Material and Methods	39
4.1.1	Dictionaries	39
4.1.2	Preprocessing of literature	41
4.1.3	Entity Annotation	41
4.1.4	Extension of Chemical Entities	43
4.1.5	Synonym Recognition	45
4.1.6	Relation Extraction	45
4.1.7	Table Processing	46
4.1.8	Quality	48
4.1.9	Curation of Relations	49
4.1.10	Biohealth Database Design	50
4.1.11	Bioactivity Data	52
4.2	Results and Discussion	53
4.2.1	Relation Extraction Tool	53
4.2.2	Curation of Text Mined Relations	55
4.2.3	Overview of Relations	56
4.2.4	Combining Bioactivity Data	58
4.2.5	Structure Activity Relationship	58
4.3	Summary	60

After proofing our concept as explained in chapter “Secondary Metabolites of Indonesian Plants” and creation of a gold set the next step is to develop a tool for automated

extraction of natural products found to be produced by biological sources. This task is called information or relation extraction. It can be divided into several subtasks. The first one is preprocessing of the publications, followed by the so called named entity recognition (NER). NER describes the classification of real-world objects. Classical named entities are “Person”, “Organization” or “Locality” which would be used for example for “Angela Merkel”, “BASF” and “Berlin”, respectively. In case of the relations targeted here, the corresponding named entities are “biological source” and “natural product”, or broader “chemical compound”. In the last step the actual relation extraction is performed. Here connections of named entities of interest are drawn and evaluated if they are relevant for the issue. In our case connections of sources and chemicals, that indicate that the chemical substance was isolated from and produced by a source should be extracted, whereas those, indicating that a source had uptaken a chemical from the environment should be ignored. The subtasks are explained in detail in the following.

4.1 Material and Methods

To extract the desired information from a text in an automated way, in our case which natural sources (plants, fungi, algae) produce which natural products (chemical compounds), several steps are necessary. First of all, there is the preprocessing of the texts of different origins from which the information should be extracted to have a standardized formatting as starting point. Afterwards the texts are divided into single sentences and those into entities (in the easiest cases words). Those entities are then tagged as for example chemistry or plant. The last step is the actual extraction of the information, here the natural products with the respective producing natural sources. The different steps are explained in detail below. The whole pipeline is programmed in python 2.7.

4.1.1 Dictionaries

To annotate single entities which their responding tags (this is described later) dictionaries are used among others. The dictionary of plant names contains latin names provided by the cooperation partner from Leipzig (University and German Biodiversity Center iDiV) and contained 11,419 species (7,689 unique species according to The Plant List⁹⁴) extracted from the flora of Java⁵⁵. Furthermore, manually a list of so called trigger words were created which indicate that a natural product was extracted from

a plant or is produced by plants (see Table A6). Plants got the tag “PL” and trigger words “ISO”. Additionally, negation words (“NEG”, e.g. “not”), stopping words (“STO”, e.g. “resistant”), words indicating that a mentioned chemical is actually an extract solvent (“EXT”, e.g. “extract”), and words of special type (“OTH”, e.g. units) were also added to the dictionary. Also, some words indicating that the word before or after could be a natural product are stored as “CHEM”, e.g. “derivate”. The treatment of

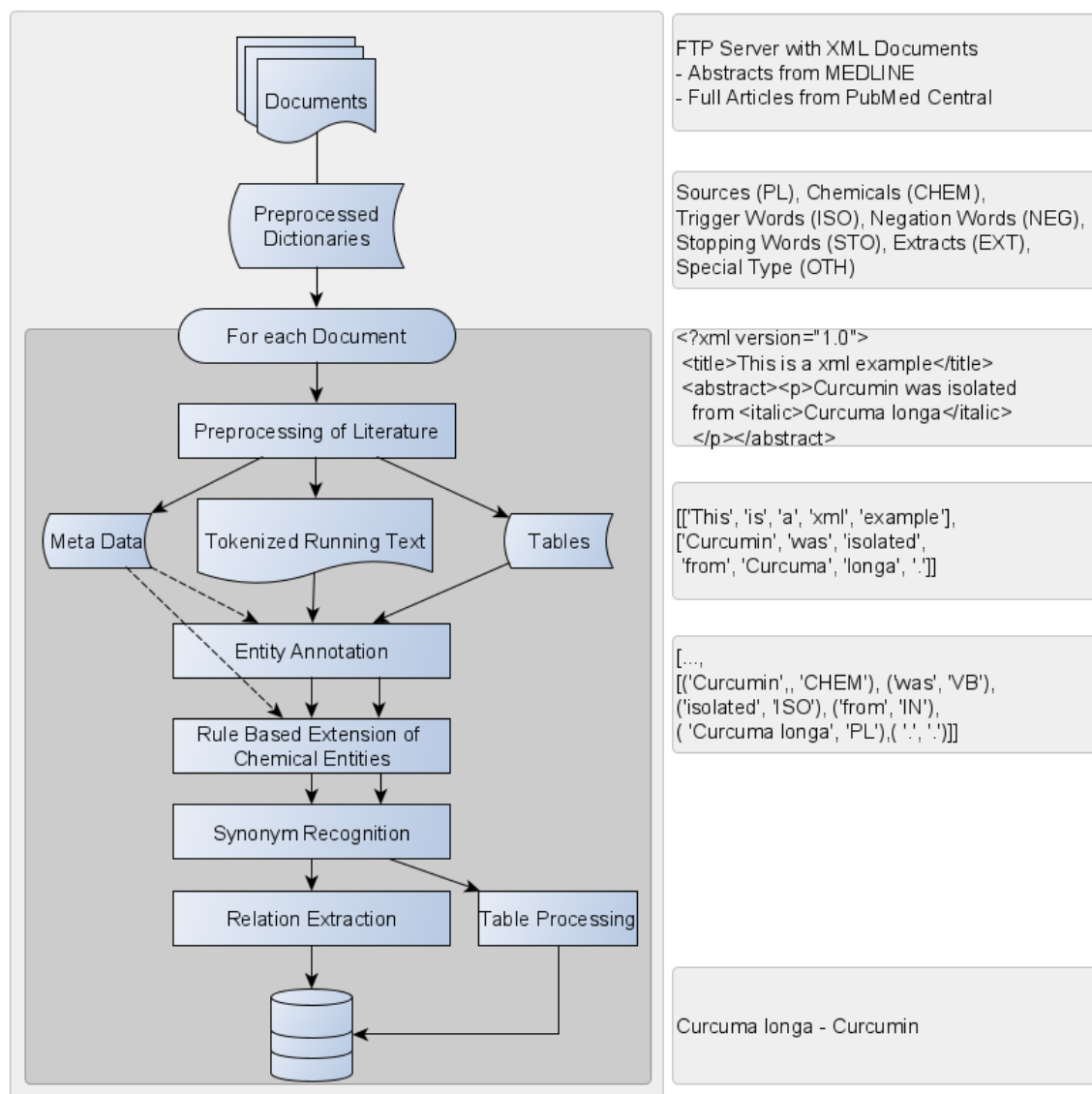


Figure 4.1: Scheme of text mining process. On the left side in light grey the initial input is shown, in dark grey the processing for each document is displayed. On the right side there is an example text snippet and how it is processed during the progress.

the entries of the dictionaries is explained in Entity annotation.

4.1.2 Preprocessing of literature

First step of the preprocessing of the literature is the actual opening/reading of the text. This step depends on the resource of the text to be mined. In this study, full text articles from PubMed Central⁹¹ (open access subset and the author manuscript collection) and MEDLINE abstracts are used and downloaded via the FTP server.

In both cases the texts are formatted in xml. The python package BeautifulSoup is used to process the xml files. From the xml tags, metadata is extracted, for example which words are written in italics, since those have a higher probability to be source names (plant species are usually put in italics, while Hybrids are plain in botanical nomenclature). Furthermore, bold numbers are marked as possible chemical labels and references within the texts extracted. The different text bodies (caption, abstract, body) and tables are extracted as basis for the further processing for which the nltk (Natural Language ToolKit)⁹⁵ package is used.

The running text is segmented into single sentences using the function PunktSentenceTokenizer. This sentence tokenizer uses a model built by an unsupervised algorithm for abbreviation words, collocations and words that start sentences. However, this approach is developed and therefore optimized for everyday language and not for scientific texts. Therefore, in the preprocessing dots in “*et al.*”, “Fig.”, “Tab.” etc. are replaced.

4.1.3 Entity Annotation

Entities are in the easiest case the singles words of a sentence like “and”. However, names like “*Aloe vera*” should also be treated as an entity. The next step of the text processing is to tokenize the single sentences into entities and to annotate those, e.g. “*Aloe vera*” as a plant species, or “Curcumin” as natural product. In the classical text mining, named entities are for example Person, Organization or Place. The step of annotating those is called Named Entity Recognition. However, we customized named entities, such as Source (in the following “PL”) or Chemical (“CHEM”). Therefore, dictionaries, rules and trained models are used to annotate the different entities round out.

The raw text without tokenized sentences is processed using the ChemDataExtractor python package⁹⁶. The function “Document” is used. This processes the text in a similar way to gain a tokenized text, however, it also contains a trained model to

recognize chemical entities in the text. It is trained to find labels, names of compounds, methods and spectra data. Since this model was trained on a different kind of texts than publications about natural products, it has problems to recognize trivial names of natural products. But it performs very well in recognizing full IUPAC (International Union of Pure and Applied Chemistry) names and order the respective labels to the names. Therefore, the recognized chemical names are used to fill a publication-specific dictionary with those as a starting point for chemical tagging. For our purposes a customized version of ChemDataExtractor is used. Labels are normally identified only by searching for numbers in the text, however, this led to several problems in our training set (wrong recognition of labels). Therefore in the preprocessing we stored the information which numbers were written bold. In the processed text “1” is replaced by “label_1” and the ChemDataExtractor is changed in the way that it does not search for number but for a number with the prefix “label_”.

Also the information of words formerly written in italics are used now to identify potential plants. Additionally to be written italic we defined rules to decide whether the italic text snip could be a plant or not. In a second step regular expressions are used to exclude snippets like “*in vivo*”:

```
> \A[A-Z]\.\s[a-z]{3,}\Z  
> \A\b[a-zA-Z]{3,}\b\Z  
> \A\b[a-zA-Z]{3,}\b\s\b[A-Z]?[a-z]{2,}\b\Z
```

The first regular expression represents a capital followed by a full stop and a word in lowercase, e. g. “*A. vera*”. The second one represents a single word written capitalized, e.g. “Aloe” and the third expression stands for two words, where the first one has to start with a capital and the second might be capital or lowercase. This includes “*Aloe vera*”, but also “Curcuma Longa”, which is an Indonesian herbal drug (not the same as “*Curcuma longa*” as plant species), would not be excluded. However, still words like “Nature Communications” or “Bioresources” would end up as potential plant names. This is why in a third step, the snippets are checked with rules for the nomenclature for plant names as it can be find at www.iapt-taxon.org/nomen/main.php (accessed 19 September 2017). If the snippet also complies to the nomenclature rules for genera and species it is stored as potential plant in a publication specific dictionary and treated in the following in the same way as the other dictionaries. Additionally, also all abbreviation of the names (“*A. vera*” for “*Aloe vera*”) are stored, and identified

abbreviations are tried to be ordered to full species names identified elsewhere (usually earlier) in the text.

The entries of the dictionaries as well as the tokenized sentences were processed with the ChemDataExtractor ChemWordTokenizer to gain the entities, since this one performs better with chemical texts than the nltk tokenizer. This step is performed with the dictionary entries since one has to know how the entries would look like in the processed text. For example “*Aloe vera*” would be processed by the ChemWordTokenizer to two entities “*Aloe*” and “*vera*”. This information is given to the nltk’s MWETokenizer, which is then used on the tokenized full text to stick there “*Aloe*” and “*vera*” again together and nltk’s UnigramTagger filled with the dictionaries information of tags to tag it with the respective tag, in this case “PL”. Furthermore, this tokenizer tags all other entities which were stored in the dictionaries.

All remaining entities which have until now no tag are labelled according to their parts of speech (lexical categories) using nltk’s pos_tag. This part-of-speech tagger uses the Penn Treebank tagset. The labels used are CC for coordinating conjunction, adverbs as RB, prepositions as IN, nouns as NN, verbs as VB, personal pronouns as PRP and adjectives as JJ. Labels of verbs and nouns can be further extended according to the tense or quantity, respectively. VBP for example stands for a verb in past tense and NNS for a common noun in plural.

Now the publication text is tokenized into sentences, those into entities and all entities are initially annotated.

4.1.4 Extension of Chemical Entities

The annotation of chemical entities is based only on some general terms in the dictionary and the result of the ChemDataExtraktor until now. However, since the ChemDataExtraktor model was built on a different kind of literature, especially common names of natural products are not recognized and therefore, they are not annotated yet. To overcome this problem a set of rules was introduced to extend the already identified chemical names in the text.

Refinement of Names: In natural product research it is common practice to name derivatives of one compound the same way with only one capital letter differing after the actual name, e.g. “Burkinabin A” and “Burkinabin C”. Consequently it is easier to

write “Compound A-D”, than name each one itself. However, the ChemDataExtractor would recognize “Compound A-D” as one chemical entity, but for our aim to find all natural products produced by a biological source it is essential to recognize each compound itself. Therefore, in this step the entity “Compound A-D” is disintegrated into “Compound A, Compound B, Compound C, Compound D” making it possible in the relation extraction to gain all four natural products.

Labels: The ChemDataExtractor already orders some chemical entities to the respective labels. If it identifies some labels, assuming e.g. label number **2**, **3** and **6**, the text is checked for the missing labels (**1**, **4**, **5**, it also checks for **7** and higher if **7** is found) and the respective chemical term. ChemDataExtractor does not recognize summaries of labels in a case like “Compound A-C (**1-3**)”. Here we already solved the summary of natural products like described in step *Refinement of Names*.

Afterwards, all labels are replaced by their full name in the text.

Enumeration Recognition: It is possible that in an enumeration of natural product names not all are recognized as chemicals and so are not tagged as “CHEM”. To spot an enumeration of chemical entities nltk’s RegexpParser to identify so called chunks of entities. One has to define a grammar describing using the tags the respective chunk one wants to extract. In the following the grammar for recognizing enumerations is shown:

```
> enumCHEM: {<CHEM>(<TYP>?<\, |\(|\)|>?<CC>?<DT>?<CHEM>)*}
> CHEMmissing: {<enumCHEM>(<\, |\(|\)|CC><\, |\(|\)|CC>?<((?!enumPL|ISO|SOLV|RESP|COL|FA|CHEMmissing).)*>)*?<enumCHEM>}
{<enumCHEM><((?!enumPL|ISO|SOLV|RESP|COL|FA|CHEMmissing).)*>
*}><enumCHEM>}
{<((?!enumPL|ISO|SOLV|RESP|COL|FA|CHEMmissing).)*><\, |\(|\)|CC>?<\, |\(|\)|CC>?<enumCHEM><\, |\(|\)|CC>?<\, |\(|\)|CC>?<((?!enumPL|ISO|SOLV|RESP|COL|FA|CHEMmissing).)*>}
}
```

The first grammar “enumCHEM” recognizes where chemical entities are connected by either a conjunction or separator like a comma and an accompanying determiner (“DT”), but also single chemical entities. The identified chunks are annotated in total as “enumCHEM”. In the second step it is checked whether between those chunks are entities enumerated which are not customized entities. The third grammar of the chunk type

“CHEMmissing” checks whether before or after “enumCHEM” non-customized entities are placed. Those entities are now annotated as “CHEM”.

Check for Typical Name-Suffixes: All non-customized entities are checked with regular expressions whether they contain a suffix which is typical for compound names (see below). If those match the entity tag is changed to “CHEM”. Of course this can introduce false positives (entities mistakenly assumed as chemical entity). However, since our very last step is the curation of all extracted natural products this chance is taken.

```
> (inin|ides|ins|ol|cin|anin|ide|onin|oids|ids|ols|cid|ate|yl|ine|ones|  
   ene|acetic|tic
```

4.1.5 Synonym Recognition

Synonyms are used for plants as well as for natural products. The ChemDataExtractor provides synonym information already, but in this step, and active search for synonyms is added. These are placed in brackets and the name before the brackets is used as main denotation. In the final emerging database the synonyms are additionally stored. If the synonym was not already tagged as “CHEM” or “PL” this is now caught up in the whole text.

4.1.6 Relation Extraction

After recognition and annotation of the entities of interest, the step of the actual relation extraction follows. For this step again nltk’s RegexpParser is used. This uses grammars to chunk texts. This means, that the entities of a sentence are checked with the grammar whether it matches or not. If the grammar matches, this part of the sentence is combined to a chunk and named how it is given in the grammar. Those so called subtrees of a sentence can be easily extracted. From those the species and chemical entities are extracted and stored as pairs of chemical and source in a PostgreSQL database described in Biohealth Database Design. In Table 4.1 the simplified regular expressions, together with text examples for the chunking are shown.

The chunking approach for the relation extraction within sentences is not transferable one to one to relations spanning over sentences. To achieve the extraction of intersentential relations the last two occurrences of “PL” and “CHEM” were stored. If then one of the following sentences a determiner (“DT” entity) is included, those are first replaced

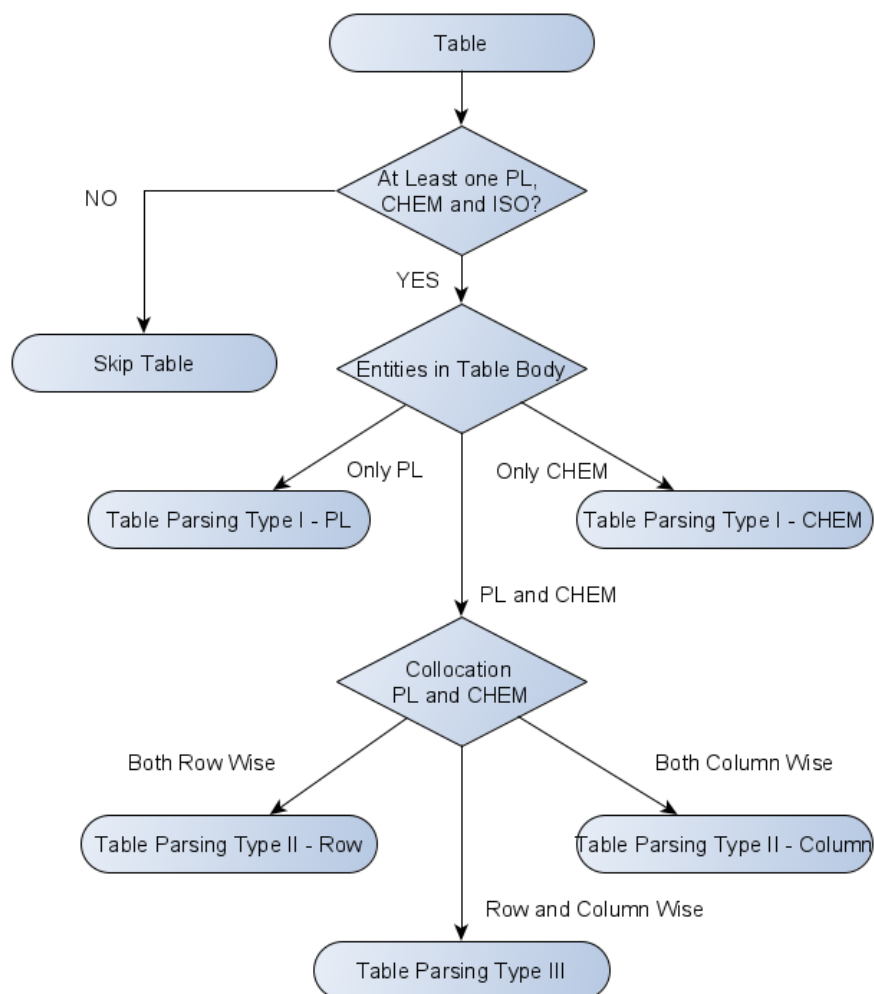
Table 4.1: Simplified regular expressions for the relation detection for “species produces natural product”-relations. NEG!.* stands for every entity beside negations. CHEM and PL can be either single entities or enumerations of those.

Regular Expression	Example Text
PL CHEM	<i>Withania somnifera</i> withanolides show anti-bacterial activity.
CHEM IN PL	Withanolides from <i>Withania somnifera</i> .
PL NEG!.* ISO NEG!.* CHEM	<i>Withania somnifera</i> produces a number of withanolides.
CHEM NEG!.* ISO NEG!.* PL	Withanolides can be isolated from <i>Withania somnifera</i> .
ISO NEG!.* CHEM NEG!.* PL	Extraction of Withanolides from the shoots of <i>Withania somnifera</i> can be achieved with ethanol.
ISO NEG!.* PL NEG!.* CHEM	The main isolation products of <i>Withania somnifera</i> are withanolides.
PL NEG!.* CHEM NEG!.* ISO	From <i>Withania somnifera</i> shoots withanolides have been isolated.
CHEM NEG!.* PL NEG!.* ISO	no example
CHEM NEG!.* COL PL	Withanolides occur in different plants: <i>Withania somnifera</i> , ...
PL NEG!.* COL CHEM	Metabolites from <i>Withania somnifera</i> are for example: withanolides, ...

by the stored “PL” and tested if one of the regular expressions matches. If not, the process is repeated with the “CHEM” entities.

4.1.7 Table Processing

As mentioned above a high amount of relations cannot be found within the text but in tables. In the *Preprocessing of literature* it was described that tables were extracted. The relation extraction from tables is solely rule based. In the following only tables were evaluated which contained at least one PL, CHEM and ISO entity, with the ISO entity is mandatory to be placed in the table caption. Four types of tables were identified which are shown in Figure 4.2.



(a) Type I - CHEM

	CHEM	CHEM
SOLV	-	+
SOLV	0	3.45
SOLV	n.d	++

(b) Type I - PL

	SOLV	dummy
PL	SOLV	dummy
PL	SOLV	dummy
PL	SOLV	dummy

(c) Type II - column wise

CHEM	PL	dummy
CHEM	PL	dummy
CHEM	PL	dummy
CHEM	PL	dummy

(d) Type II - row wise

	PL	PL
CHEM	CHEM	CHEM
dummy	dummy	dummy
dummy	4.7	5.8

(e) Type III

	PL	PL
CHEM	8.7	dummy
CHEM	1.5	dummy
CHEM	4.7	dummy

(f) Type III

	CHEM	CHEM
PL	3.7	dummy
PL	4.6	dummy
PL	2.3	dummy

Figure 4.2: Scheme of table processing. First the caption and whole table is checked for at least one PL, CHEM and ISO entity. If this is true, the caption is processed separately. Four table types which were checked for "species produces natural product" relations.

The table information is converted into a temporary SQLite database to preserve the information of row, column and therefore, cell positions. Furthermore, the search for entities can be performed easily by SQL queries. If only one PL entity or CHEM entity is mentioned, all entities of the other type is ordered to the first one. An example could be a table with the caption “Natural Products extracted from *Curcuma longa*”. Here we have only one PL entity in the caption - *Curcuma longa* - and in the table only CHEM entities and entities of other types are mentioned. Consequently, all CHEM entities are set in relation to *Curcuma longa* as “species produces natural product”.

The more complicated cases are those where CHEM and PL are mentioned side by side in case of a “species produces natural product” relation, for example in an enumeration of investigated plants with identified natural products. In this case the first step is to recognize whether it is necessary to combine information row by row or column by column. This is done by checking if one or multiple PL / CHEM can be found per row. If only one is found per row it is assumed that the pairs are ordered rowwise. Those entities can be also enumerations (“*Curcuma longa* (Zingiberaceae)” would be initially recognized as enumeration), however, the quantity refers to the number of cells with the respective entity type.

As for the relation extraction from the running text, all relations are stored in the database described in detail in *Biohealth Database Design*.

4.1.8 Quality

The performance of the text mining tool was always controlled by measuring the precision, recall and F-Score during the development. Therefore, the goldset described in Creation of a Species - Natural Product Gold Set is used. The relations extracted via the developed tool were compared to those from the goldset and categorized as true/false positives/negatives. Recall reflects, how many of all gold set relations were found and precision is the amount of the correctly identified relations out of all identified relations. The F-Score considers precision and recall. The equations can be found below, where tp stands for true positive, fp for false positive, tn for true negative and fn for false negative.

$$Precision = tp/(tp + fp) \quad (3)$$

$$Recall = tp/(tp + fn) \quad (4)$$

$$F-Score = 2*((Precision*Recall)/(Precision+Recall)) = 2*tp/(2*tp+fp+fn) \quad (5)$$

While the development of the tool those parameters are regularly calculated based on the training set. All relations were only taken into account once, although they occur multiple times in a text. The development of the named entity recognition as well as the relation extraction was performed in an iterative way always controlled by evaluating precision, recall and F-Score. It is possible to check which entities or relations were not correctly identified in the training set. Missing or incorrect entities/reactions are not evaluated manually with the test set. Therefore, the test set acts as an independent quality measurement to prevent over fitting to the training set. This means, that the parameters are evaluated and it is ensured that they do not differ widely to those measured with the training set.

4.1.9 Curation of Relations

As mentioned above the natural sources and products derived from the text mining approach have to be curated to identify errors, duplicates and to find the structures of the natural products.

Natural Sources: The identified sources, if not already contained in the dictionaries, were checked against the GBIF taxonomy backbone⁹⁷. All unidentified source names were compared to lists of fungi and bacteria (from www.mycobank.org and www.bacterio.net, accessed 21 November, 2017). However, this approach is susceptible to misspellings and typos. The global names resolver was therefore used to identify and resolve such⁹⁸. This was performed automatically for all source names which could not be found in the GBIF backbone taxonomy using the `r` package `taxize`⁹⁹.

Natural Products: Natural products were checked against several databases (PubChem, ChemSpider, ChEMBL, Chemical Identifier Resolver) either online or in a local version of the databases. Additionally, the Chemical Translation Service was used as well as `opsin`, which assigns structures to correct IUPAC names. There are also chemical

group terms (e.g. terpenoids, flavonoids) in the extracted natural products. To annotate those groups the chemical groups which are used by the tool classyFire¹⁰⁰ were used.

4.1.10 Biohealth Database Design

The extracted relations between natural products and sources are stored in a PostgreSQL database in table np_relations (see Figure 4.3). Additionally, the ID of the literature, the quantity of the relation in the text and whether the source was identified as probable herbdrug or not are stored in the same table.

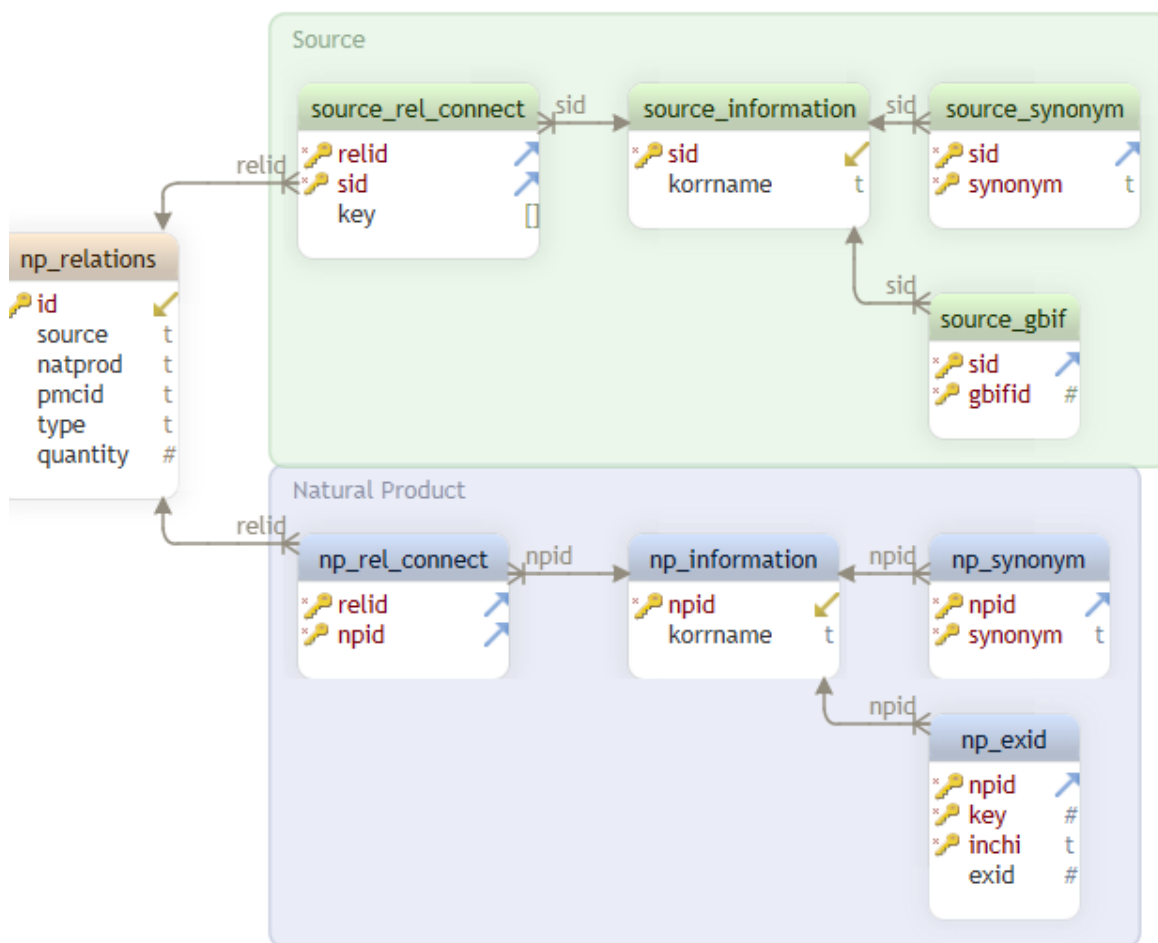


Figure 4.3: Scheme of the database containing the extracted "species produces natural product" relations together with additional information.

Table 4.2: Resources used for natural product curation and InChI source with the keys used in the Biohealth Database table np_exid. “No NP” are for example numbers only or other substantives, which were identified as natural product falsely. “Chemical classes without structure” are for example general chemical terms such as “Compound”, while for “Chemical classes with structure” a backbone structure can be assigned (e.g. “flavonoids”). “Peptides” are peptides or even proteins. The remaining classes are either databases of chemical compounds or tools. The respective reference is given in the table.

Natural Product Matching		
Resource	Database Key	Reference
No NP	0	
PubChem local	1	101
Chemical Translation Service	2	102
PubChem online	3	via python package CIRpy ¹⁰³
ChemSpider	4	via python package CIRpy ¹⁰³
Chemical Identifier Resover	5	via python package CIRpy ^{103,104}
Opsin	6	via python package Cinfony ^{105,106}
ChEMBL	7	107,108
Chemical classes	8	Classes derived from ClassyFire ¹⁰⁰
NPASS	9	27
KNAPSAcK	10	25
Peptide	11	

Information gained during the curation process is stored in further tables. All sources and natural products got an unique ID during the curation process called sid and npid, respectively. Sources and natural products which occur with different names in the extracted relations are here already deduplicated and one main name (“korrname”) is stored together with the IDs in the tables source_information and np_information. The identified synonyms are stored in the tables source_synonym and np_synonym with the respective IDs. Source_gbif connects the deduplicated sources of the relations with the GBIF database using the GBIF-ID (gbifid). The GBIF-IDs are used from a local GBIF database copy from April 1, 2019. Np_exid connects the npid with an

InChI string as structure representation and a key coding the origin of the InChI string. Additionally, the external ID, if available, is stored - for example the PubChem-ID if the natural product was identified in the PubChem database.

4.1.11 Bioactivity Data

Although the activities were not extracted from the publications it would be great to combine this information to the natural products. Therefore, ChEMBL Version 24.1¹⁰⁸ was used to gain bioactivity information. Since ChEMBL contains huge amounts of bioactivity data, here an example of an analysis is presented. All natural products were queried against a local version of the ChEMBL database using the structural information. From the resulting targets and assays, one High-Throughput Screening (HTS) Assay (PubChem Bioassay AID 602399) was chosen for further Structure-Activity Relationship (SAR) Analyses. The bioassay is a high-throughput assay testing for activity against a probable nicotinate-nucleotide adenylyltransferase (NMNAT family) encoded on the so called nadD gene (Uniprot Entry P65502 – NADDD_STAAN) and one key enzyme in the NAD(+) synthesis.

The dataset contains information for 362 388 PubChem compounds (2 821 are active). From this dataset, a test and a training set was created using 300 active and 300 inactive compounds, randomly chosen each. For all of those with MOE descriptors were calculated (all 2D descriptors, except those including van der Waals volumes) and with Python rdkit morgan fingerprints (radius = 3).

Methods used for creating a binary (only information active/inactive) structure activity relationship models (standard settings were used, adaptations are indicated in brackets) 0. MOE⁶³ pharmacophore elucidator 1. MOE⁶³ binary SAR (modelled by two different persons (1.1 and 1.2) and using MOE⁶³ contingency tool for identifying relevant descriptors (1.3)) 2. MOE⁶³ classification tree 3. Python scikit¹⁰⁹ partial least square, modified to a discriminant variant (PLS-DA) (inactive was converted to -1, active to 1, threshold = 0) 4. Python scikit¹⁰⁹ support vector machine (SVM) (linear/poly degree = 1) 5. Python scikit¹⁰⁹ random forest (n_est = 200) 6. Python keras¹¹⁰ deep neural network (DNN) (layers = 2)

The MOE 2D descriptors were also used for the Python approaches. Within these approaches additionally Morgan fingerprints calculated using rdkit¹¹¹ in Python were used as structural representation.

4.2 Results and Discussion

4.2.1 Relation Extraction Tool

As mentioned in *Material and Methods* the relation extraction consists of several steps. The annotation of natural products and sources, as well the actual relation extraction can be evaluated by measuring precision, recall and F-Score (see *Quality*).

During the development of the relation extraction tool several times precision, recall and F-Score were calculated for the trainings set. Table 4.3 shows the final values for these measurements for training and test set. This tool was trained for a high precision rather than a high recall, since the aim was to get reliable information.

Table 4.3: Precision, Recall and F-Score of the finalized relation extraction tool measured on the training and test set in %.

	Precision	Recall	F-Score
Training Set	77.3	18	29.3
Test Set	66.9	21.9	23.3

However, the tool seems to be slightly overtrained, because the precision in the test set is just around the half of the value of the training set. But one has to mention, that the test set also contains only one third of relations found in the training set.

The following query was used to extract relevant literature from PubMed Central.

```
("Constituents" OR "Chemical study" OR "phytochemical study"  
OR "Phytochemical investigation" OR "chemical investigation"  
OR "Isolation" OR "structure elucidation" OR "Chemical composition"  
OR "new compounds" OR "novel compounds") NOT ("Genomic" OR  
"Proteomic" OR "Transcriptomic" OR "DNA")
```

This resulted in 153,444 open access and 58,783 manuscripts. PubMed and MEDLINE provide their collections for download. All available (by August 2017) MEDLINE abstracts were used as input for the developed relation extraction tool. The PubMed collection was filtered with the shown query, however, not all PMCID's were included in the downloaded article set.

In total 247,269 relations from 25,298 PMC articles and 82,068 MEDLINE abstracts

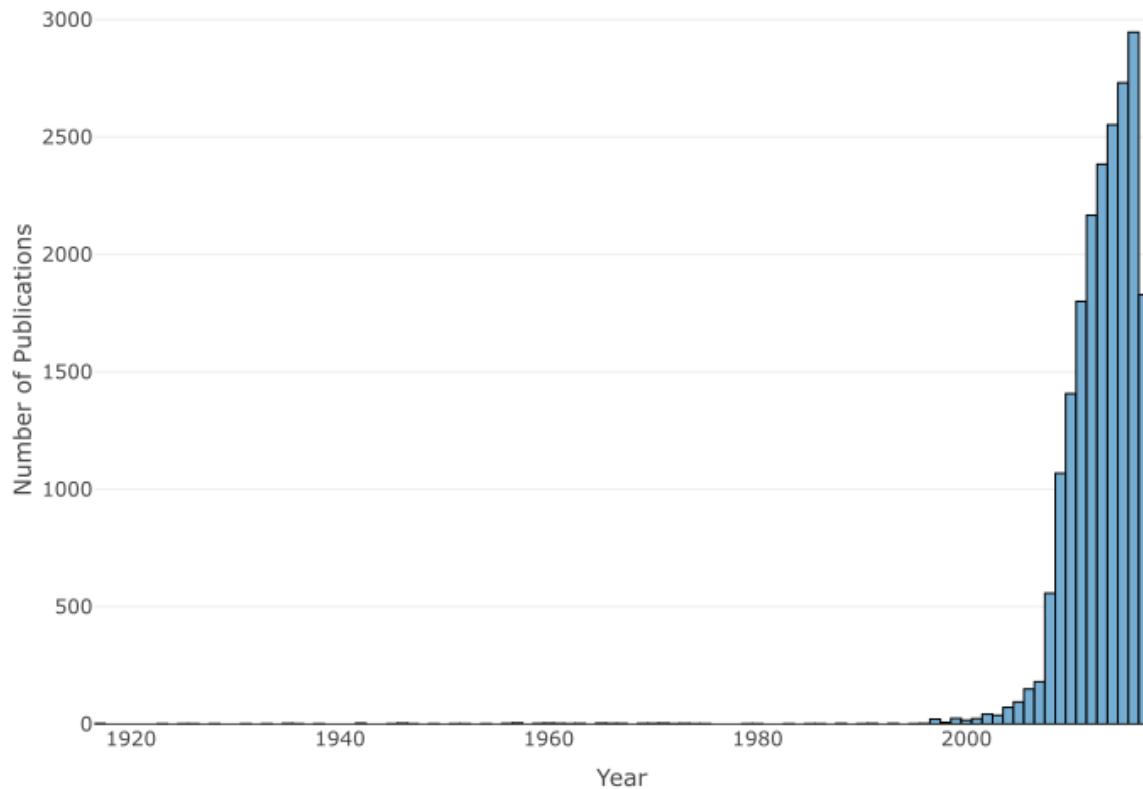


Figure 4.4: Number of PubMed Central full text publications used for the text mining per year of publication.

(intercept of 1,886 abstracts). Those consist of 28,441 sources and 117,981 natural products. The detailed distribution of the different literature sources can be found in Table 4.4. Some entities or relations were found already in the abstracts of the PMC articles (Abstract Intercept), while the overall intercept derives from the comparison of all relations gained per text base.

Figure 4.4 shows that most of processed publications were published after 2005 since then most publications are available digitally, while only a minor part published earlier is digitalized. This was a major drawback of the automated approach. There are tools for digitizing scanned documents, but there are often errors in the generation of the texts, especially if they are in the style with two columns, for example. This is an advantage of manual approaches, such as those used to fill the DNP²¹. However, it can also be seen that the number of publications is steadily increasing per year. Although Chassagne *et al.*¹¹² used a manual approach the trend in the number of publication was

also observed by them. To quickly screen this large number of publications for relations is the advantage of the automated process. Unfortunately, unlike as in some other areas of science (e.g. molecular biology) in natural product research older and even very old papers are of equal value to and indispensable for current research.

Table 4.4: Number of text bases and entities for PubMed Central and MEDLINE, together with the overall intercept or intercept in the matched abstracts.

	Text source	#Relations	#NPs	#Sources
PMC	23,412	115,831	41,084	19,011
MEDLINE	80,182	106,704	51,087	2,838
Overall Intercept	NA	6,211	7,001	1,343
Abstract Intercept	1,886	3,821	2,309	903

4.2.2 Curation of Text Mined Relations

To extract false positive entities, duplicates and synonyms from the result set, sources as well as natural products were checked independently with different databases. It was not necessary to correct the natural sources derived from the MEDLINE abstracts, because in those we can identify the natural sources only by using a dictionary of

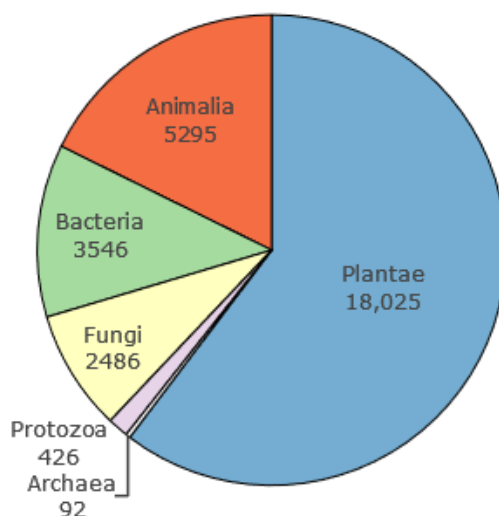


Figure 4.5: Number of structurally unique natural products per kingdom identified via GBIF taxonomy.

Table 4.5: Number of unique natural products identified with each database or tool.

	Identified natural products
PubChem (local version)	8,468
PubChem (online)	5,134
ChemSpider	1,452
ChEMBL (version 24, local)	1,005
Chemical Identifier Resolver	2,230
Chemical Translation Server	26,762
Opsin	1,996
Error identified	1,062

species names and this was based on the list of Indonesian plant names. The curation of the natural sources extracted from the PMC full articles showed that 92 % of the extracted natural sources are correct plant, fungi or bacterial species, genera or families (identified via the GBIF backbone taxonomy⁹⁷, while 0.5 % are herbal drugs. In 0.5 %, 0.8 % and 3.2 % of the cases the sources were synonyms, had typos in the name, or are abbreviations, respectively. The remaining 3 % are not existing species (1 % errors by the authors, 2 % by the text mining tool). It was possible to assign structures and names to 56,200 of the 117,981 natural products, corresponding to 40,926 unique (compared via InChI) natural products. How many natural products were identified with each database or tool can be seen in Table 4.5.

Already some errors were possible to be identified. Those are at the moment either natural sources names or proteins which were incorrectly annotated as chemicals by the relation extraction tool or labels for true natural products which were not resolved by the text mining tool.

4.2.3 Overview of Relations

For the overview, only those relations where source and natural product were confirmed as described above were used. In Figure 4.5 one can see that natural products from all kingdoms were identified. The fact that the number of natural products is highest for plants (Figure 4.5) is not surprising since a) research on natural products is historically focused on plants (and more recently microorganisms), and b) in addition, the text mining tool has been trained mainly for plants and identifies them with higher probability

than species of the other kingdoms. The review of the DNP²¹ by Chassagne *et al.*¹¹³ shows 67 % of natural product belonging to the plant kingdom. In our dataset 70 % of natural products belong to plants. In both analyses, fungi show the same ratio (10 %), while our approach identified more natural products from Animalia (20 % vs. 13 %).

The obtained structures in InChI string representation were used to classify the natural products according to their chemical substructures. Therefore, the defined SMART Table A2 in MOE were used. Combining this information with the GBIF backbone taxonomy one can see how the different chemical substructures are distributed over the different kingdoms.

Figure 4.6 shows how the chemical substructures identified via SMARTS in MOE are distributed over the kingdoms. The figure shows relative ratios since one natural product can have several chemical substructures: A steroid will be classified as a terpene, or a glycosylated flavonoid as flavonoid as well as a sugar. Despite this, one can see that

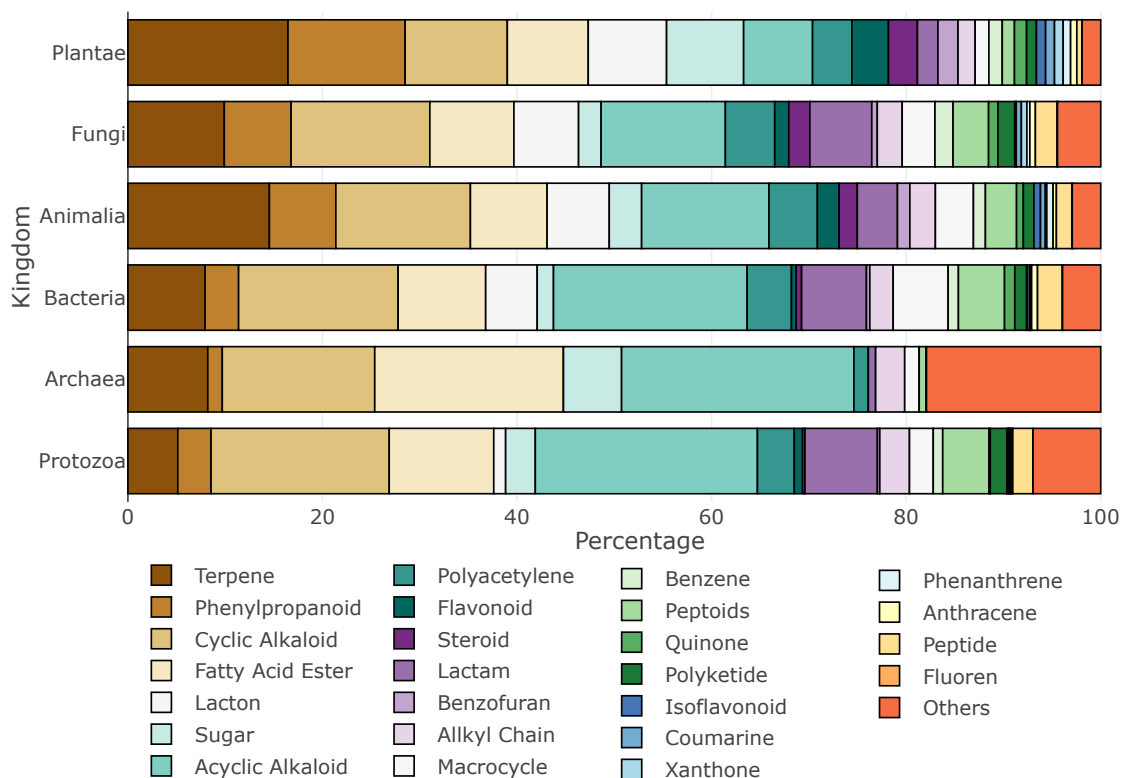


Figure 4.6: Ratio of chemical substructures identified by SMARTS in MOE of natural products per GBIF taxonomy kingdom.

terpenes, phenylpropanoids, cyclic alkaloids and fatty acid esters represent a large proportion of natural products in all kingdoms. Archae show the highest number of natural product which remain unclassified with the SMARTS approach. However, the total number of natural products identified for Archae is only 92 in this dataset anyway.

4.2.4 Combining Bioactivity Data

In tables Table A7 and Table A8 the assays, respectively targets of ChEMBL which had the highest matches of natural products queried against the activity information. For further analyses the target *Staphylococcus aureus* was chosen because of the significance of multiresistant strains against antibiotics of this organism. Focusing on this target an enzyme (nicotinate-nucleotide adenyltransferase) assay was chosen for which high throughput data are available.

4.2.5 Structure Activity Relationship

Except the pharmacophore elucidator, all methods resulted in acceptable models (see Table 4.6). The pharmacophore elucidator did not work, because the structure of the compounds is quite different. There are a lot of molecules that seem to mimic the substrate/transition state, however, there are also quite large ones which seem to inhibit exclusively or additively the protein-protein interaction (the target enzyme is a dimer).

Table 4.6: Recall, Precision and F-Score values for identifying the active molecules in the test set (1.1 and 1.2 two different MOE⁶³ SAR models, 1.3 MOE⁶³ contingency tool, 2 MOE⁶³ classification tree, 3 scikit¹⁰⁹ PLS-DA, 4 scikit¹⁰⁹ SVM, 5 scikit¹⁰⁹ random forest, 6 keras¹¹⁰ deep neural network).

Method	Recall	Precision	F-Score
1.1 (SAR)	74	74	74
1.2 (SAR)	75	80	77
1.3 (Contigency Tool)	78	81	79
2 (Classification Tree)	78	86	82
3 (PLS-DA)	71	72	71
4 (SVM)	84	85	84
5 (Random Forest)	88	85	86
6 (Deep Neural Network)	83	86	84

The learning approaches (4, 5, 6) clearly outperform classical classification methods. Although, the precision of MOE classification tree is also quite high (comparable to a random forest approach). This is again mainly due to the fact of the quite diverse structures of compounds. The learning approaches can, to a certain content, build models recognizing different structurally similar groups of compounds. In contrast, classical approaches like binary SAR (in MOE based on PCA), and PLS-DA only differentiate in active and inactive group. With large structural differences, these analysis methods cannot adequately distinguish the two groups.

The resulting models of the learning approaches were used to predict the activity of the natural products for which structural information is available. Figure 4.7 shows that 4,627 natural products were predicted as potentially active by all three models. These were reviewed manually and appear plausible. There are molecules that could block

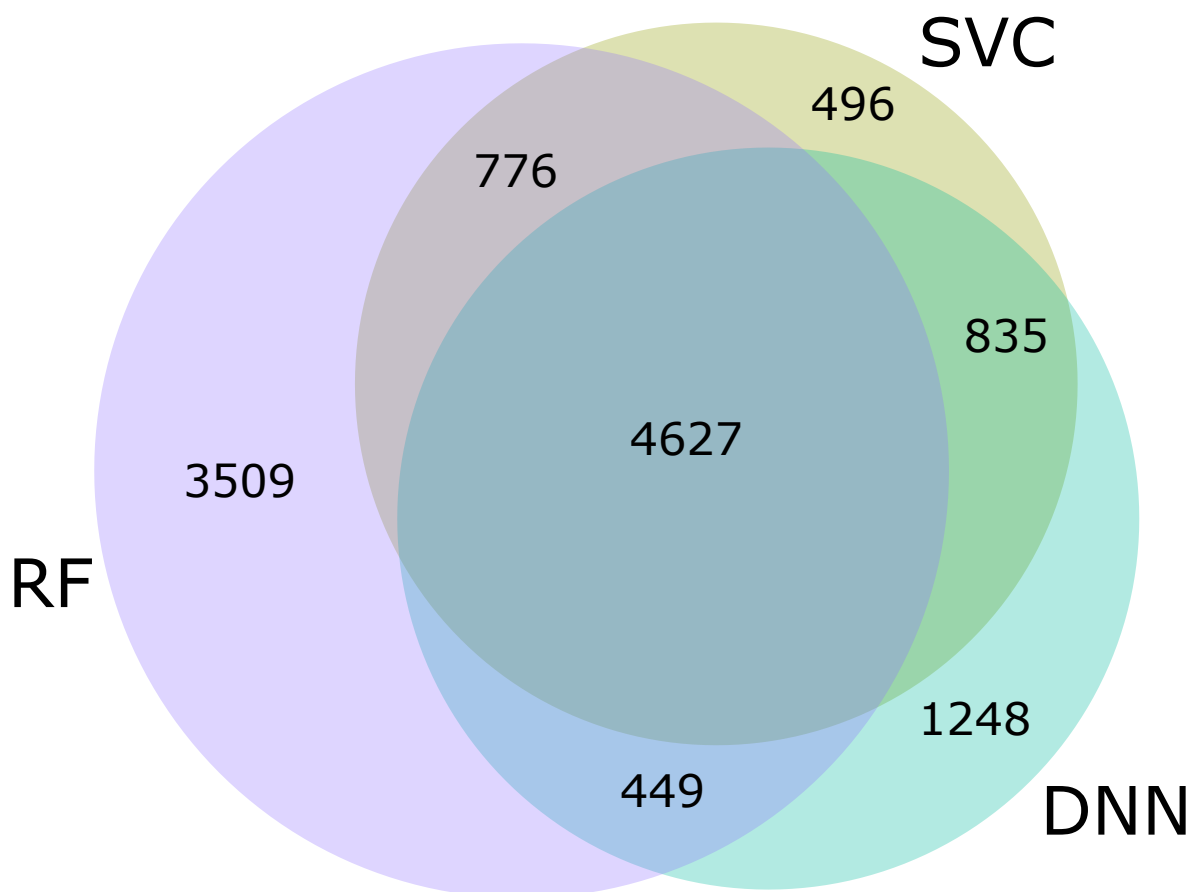


Figure 4.7: Venn diagramm of prediction of the random forest (RF), support vector machine (SVM) and deep neural network (DNN).

the active pocket as well as the protein-protein interaction, just like the underlying active molecules of the HTS assay data. Checking with the PubChem-IDs indeed 745 of the identified molecules are already part of the assay data. The number could be even bigger, since the PubChem-IDs were used for comparison. However, from these 745 only 18 are really active in the assay. 2.4 % seems to be a very low ratio of correct prediction in comparison to the precision and recall of the test set, but it is higher than the ratio of activities of the HTS data (0.8 %). However, the chemical space used should be checked if it covers sufficient natural products in the HTS data used for training. The chemical space describes the (parameters of) structures of compounds in the data sets. If the chemical space of the natural products is different from that in the HTS data applied in the models this would explain the low correctness of prediction. Indeed natural products are known to cover a chemical space, very different from synthetic or general chemical space.¹¹⁵ If the chemical spaces are similar, it would show that the models were mistrained. However, since only 300 compounds were used for training (due to performance limits of MOE) and since such differences are known, the first scenario is more likely. So the result shows that it is a beginning, but the models have to be trained on more data, and on better suitable data (which are available).

4.3 Summary

A relation extraction tool was developed as a python package, where the single steps article preparation, named entity recognition, and parsing can be used independently. Using this tool, in total 117,981 natural products could be found in MEDLINE Abstracts and PubMed full articles for 28,441 natural sources with 247,269 relations. The curation of natural sources showed that indeed 97 % are sources and mainly plants. Therefore, the tool achieved higher precision in recognizing plants in the application than in the development. In the case of the natural products for 48 % structures were found and therefore those are actual chemical names. This is less than the precision determined via the goldset. However, the reason for this approach is the fact that the chemical databases are by far not complete. The extraction of new natural products is ongoing and not all newly identified chemical compounds are added to the databases, or the trivial names are missing. Checking the natural products not identified within the here described curation are Antrodol B, Flavichalasin B, or Iritectol A as examples for classical natural product names, but also 2-*O*-deacetylorthosiphonol J, 2-hydroxy-11-ketonorzoanthamide B or 4-*O*-methylberberammonium C as examples for natural product names mixed with IU-

PAC nomenclature. Furthermore, the names not found in the curation step are chemical class names mixed with IUPAC nomenclature (e.g. 4,5,7-trihydroxy flavonoids), names including “derivative” (e.g. 5,6-Dihydro- α -pyrone derivatives), non-solved abbreviations or labels, or obviously malidentified chemical entities. In consequence, one can assume, that the percentage of correctly identified natural products is higher than the given 48 %. Possible ways to check the unconfirmed natural products would be to repeat the curation process after a while to cover names which are added later, to add further databases to the curation step (for example spectral databases), to confirm manually or to perform structure recognition within the manuscripts. Optical structure recognition tools have problems, especially with bridges in molecular structures (which are unavoidable in cage structures), long bonds, large molecules, unknown bonds, overlaps, annotations in structures, salts, and chirality annotations.¹¹⁶ Nevertheless, the freely available tool OSRA¹¹⁷ was tested with one publication¹¹⁸ on natural products. But there was no structure correctly solved, and accordingly optical structure recognition was discarded for the extraction tool described here.

A possibility to enhance precision and recall for named entity recognition and relation extraction could be machine learning. Using the python package spaCy¹¹⁹ first tests were performed whether the named entity recognition could be improved with machine learning. Indeed the precision of the chemical recognition could be extended to 86 % with a corresponding recall of 67 % in the test set. For the recognition of entities of type “PL” there was no remarkable improvement. SpaCy uses the surrounding words for the recognition of entities, which is beneficial for recognizing chemical names. Of course, also plant names may be surrounded by similar words, but due to the structured way of naming they can be already identified via rules with a high precision. Bioactivity data could be gained by either adding another relation (natural product - bioactivity) extraction to the text mining tool, or by connecting the natural products to existing information. The second approach was performed with ChEMBL assay and target information. For a number of natural products, activity data are available. However, assays and targets are not classified in a way we used for example in Secondary Metabolites of Indonesian Plants (e.g. anti-bacterial, anti-viral). Therefore, a classification must be performed, or, as we showed with the SAR analysis, not only the information of the natural products for this assay, but the whole assay data is used and afterwards applied to the remaining natural products.

5 A World Wide Overview of Plant Natural Products

5.1	Material and Methods	63
5.1.1	Data acquisition	63
5.1.2	Classification of natural products	63
5.1.3	Global distribution of plants and natural products	63
5.1.4	Habitat analysis	64
5.2	Results and Discussion	64
5.2.1	Primary open access publications as source of information	64
5.2.2	Which are the most common natural product classes and how are they distributed?	65
5.2.3	Are natural products known from all over the phylogenetic tree of plants?	65
5.2.4	Are there less studied regions?	68
5.2.5	Are species-rich tropical rainforests really the source of most natural products?	70
5.2.6	Do chemical and biological diversity go hand in hand?	70
5.3	Summary	72

In chapter Secondary Metabolites of Indonesian Plants it was shown that combining phylogenetic, taxonomic, geographic, and chemoinformatic data on natural products and the producing natural resources is a promising method to obtain more comprehensive knowledge and identify new targets for natural products research. Now, similar analyses will be applied to the data obtained in chapter Creation of a Global Data Set. Again,

plants will be selected as organisms because for them the largest number of natural products is available in the resulting database.

5.1 Material and Methods

5.1.1 Data acquisition

The data used for the following analysis were acquired as described in Creation of a Global Data Set. From this set 5,628 plants (genus, species or subspecies) with occurrence information from GBIF⁷¹ were selected. For these 16,199 secondary metabolites are present in 50,817 relations in the database.

5.1.2 Classification of natural products

Using the InChI string representation of the natural products the ClassyFire tool, which classifies structures using the ChemOnt ontology¹⁰⁰ was accessed via an R script kindly provided by H. Treutler (at this time Leibniz Institute of Plant Biochemistry). Only the primary classification (not the alternative classifications) was used for the evaluation. For each node in the ontology the number of matched natural products were counted and afterwards all nodes with more than 100 matches were illustrated (see Figure A1). The size of the nodes corresponds to the number of matched natural products. The ratio for each category (number matches / total number of natural products) are given in Table A9.

5.1.3 Global distribution of plants and natural products

The world map was rasterized with a resolution of 0.833 arcseconds. All available plant occurrences were downloaded from GBIF⁷¹ and checked for each grid cell in the raster how many and which plants are located within. The amount of unique plants (genus, species or subspecies) per grid cell is plotted in Figure 5.2A. Since in the data acquisition step all plants identified in the text mining approach were mapped onto the GBIF backbone taxonomy⁹⁷ from the former analysis it was possible to also count the unique text mined plants per grid cell, visualized in Figure 5.2B. Then for each grid cell all metabolites for the respective plants were identified and also counted uniquely (Figure 5.2C). In Figure 5.2D than the number of text mined plants versus the number of metabolites are visualized. The relation between the number of plants and the

number of metabolites were checked for correlation two different regression methods: the linear regression and by fitting a growth function using the following equation:

$$M(p) = S - (S - M_0)e^{-kp} \quad (6)$$

with M being the number of metabolites and p being the number of plants) Both regression showed significant correlation between the two measured values, however it was not possible to discard one.

5.1.4 Habitat analysis

For the analysis of how many of the GBIF and text mined plants grow in each habitat again the occurrence information of GBIF was used along with the global terrestrial ecoregions map published by WWF⁷². From this we were able to conclude which plants grow in which habitat and again the number of plants were counted uniquely for each habitat.

5.2 Results and Discussion

5.2.1 Primary open access publications as source of information

Since most natural products, once isolated from a source are published, the primary literature is a resource for a comprehensive information overview. PubMed Central¹²⁰ and MEDLINE¹²¹ are resources for freely available literature. The first provides 5.5 million free full text articles, while the latter contains 29 million entries, 85 % of those being abstracts. Those were used as basis for a relation extraction approach to gain natural product names along with their biological source from the publications. The presented studies base on 5,628 plants (genus, species or subspecies) for which we know 16,199 secondary metabolites connected by 50,817 relations. We know that this is far beyond the number of natural products that was already isolated from plants, however, since this is an unbiased collection of the knowledge the community has about natural products, we think that this can provide information of general statistical relevance and that thus we can address some of the questions introduced in the beginning with this data set.

5.2.2 Which are the most common natural product classes and how are they distributed?

To have an idea how the natural products studied here are composed all natural products were classified using the ChemOnt ontology based ClassyFire tool¹⁰⁰. ClassyFire uses a structure-based (SMARTS - SMILES arbitrary target specification) chemical taxonomy (ChemOnt). It consists of over 4,800 different categories which are linked to existing databases and ontologies such as ChEBI (Chemical Entities of Biological Interest)¹²² or MeSH (Medical Subject Headings)¹²³. In a taxonomic way the natural products are classified from coarse classes to finer classes. The service of ClassyFire was accessed via an R script kindly provided by Hendrik Treutler. In an iterative way it sends bundles of structures (e.g. 100) of the total number to the ClassyFire server and collects the returned classification of the compounds.

Most of the natural products are lipids and lipid-like molecules (5,511 matches – 33 %) (see Figure A1) followed by prenyl lipids (called prenol lipids in ClassyFire) (3,790 – 23 %), phenylpropanoids and polyketides (3,030 – 18 %) and organoheterocyclic compounds (2,629 – 16 %) as core structures, where the prenyl lipids are children of the lipids and lipid-like molecules. 5 % of the here studied natural products are alkaloids and derivatives, 7 % flavonoids and 6 % steroids and steroid derivatives. A detailed list of the ChemOnt classes which matched at least 100 natural products can be found in the supplementary material (see Table A9).

5.2.3 Are natural products known from all over the phylogenetic tree of plants?

Our proprietary *in-house* tool to extract the desired information identifies natural product source names in a rule based approach using the rules stated in the international code of nomenclature for algae, fungi, and plants (Melbourne code)¹²⁴. Therefore, we achieved an unbiased recognition of plant names in comparison to approaches that use for example a dictionary of different plant names.

To check on the one hand if the sampled information is really unbiased in terms of plant genera and on the other hand to identify regions of a phylogenetic tree, which are over- or under-represented in natural product research, the plant genera were aligned with a phylogenetic tree of land plants. For this the phylogenetic tree constructed by Hinchliff and Smith¹²⁵ comprising around 80 % of plant diversity was used. For the

13,093 genera of the tree we had natural product information for 1,983 genera which corresponds to 15 % of the land plants. This is in agreement to the studies of Dias *et al.*¹²⁶ and Cragg *et al.*¹²⁷ who also state that only 5-15 % of biodiversity was evaluated for their natural products.

Using a phylogenetic clustering approach ('nodesig' algorithm¹²⁸) we identified regions of the phylogenetic tree which are significantly less or more studied than one would expect from the average number of studied genera (see Figure 5.1). With this approach we found 287 plant families which seem to be over- and 84 families that are under-proportionally studied. Interestingly, for 307 families we got a signal for both, over- and

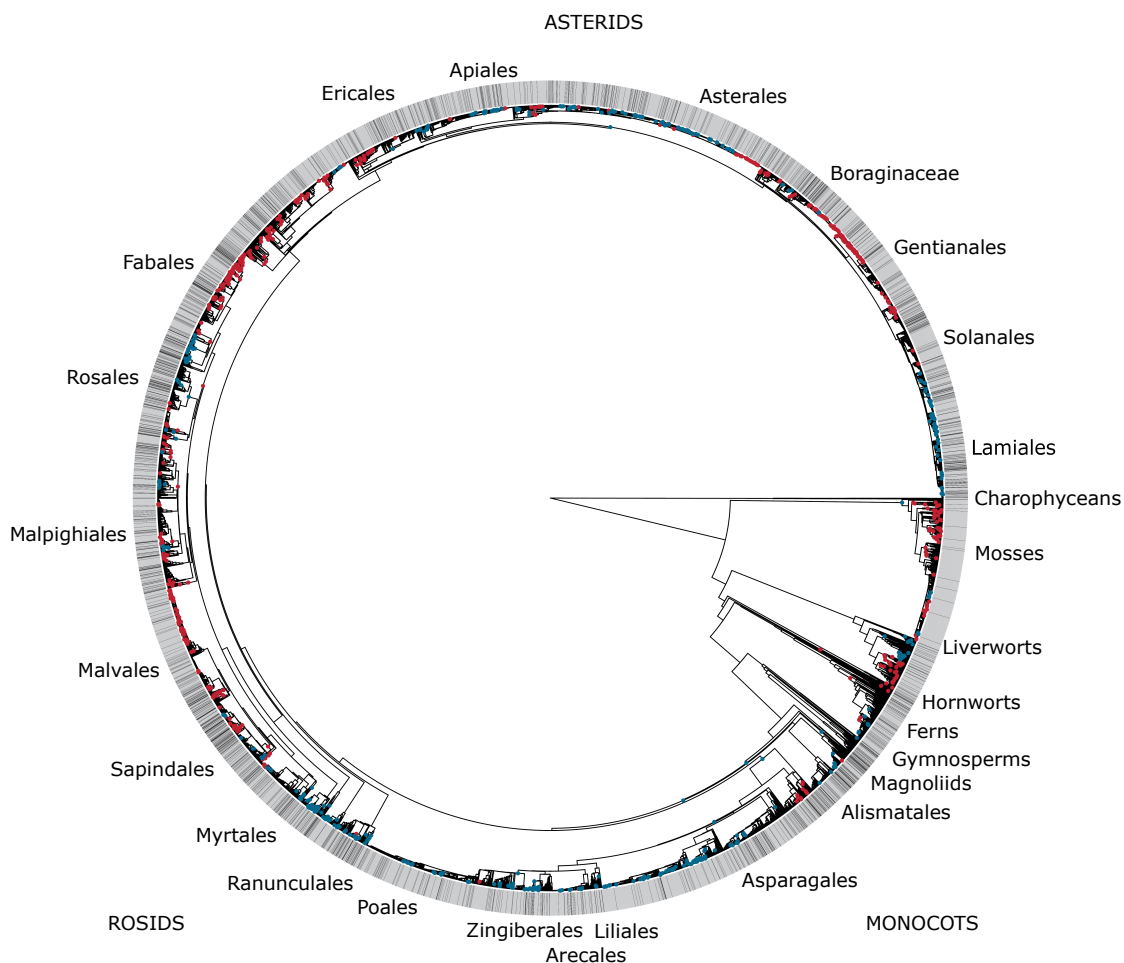


Figure 5.1: Phylogenetic tree of plants. The black lines indicate which genera were identified in the text mining approach. Red and blue spots indicate which node children show plants with significantly higher or lower natural product information, respectively.

under-representation. That can happen since in a phylogenetic tree between the tips (genera) and the family nodes there are several nodes were the tree branches. Those nodes in between can then show different signals dividing families in parts of over- and under-studied regions. Examples for well-studied plant families according to this investigations are for example Cannabaceae, Hypericaceae, Lauraceae or Piperaceae, which is not surprising, since all are families of prominent plants used for example in traditional medicine and therefore were studied in detail for their metabolite composition. Additionally, several species of these are easily accessible, even commercially grown, and they are distributed globally. More interesting in terms of the discovery of unknown natural products are of course plant families that are less studied. In this category fall only plant families of the classes Andreaopsida, Bryopsida and Sphagnopsida which are mosses (Bryophyta) and of the class Haplomitriopsida, which is a liverwort (Marchantiophyta). Although mosses and liverworts are rich sources of metabolites and can easily be cultured they were and are neglected in natural product research.^{129–132}

This is mirrored on the one hand in being classified as under proportionally studied, on the other hand parts of the Bryophyta and Marchantiophyta are also classified as over proportionally studied (for some subgroups or species, e.g. model plants in biology) suggesting in combination with the existing knowledge of metabolites from mosses, that especially they might be a valuable source of until now unknown natural products.

In contrast to the mosses, the families which show both signals are at least partially chemically investigated. Examples for these are Fabaceae, Lamiaceae, Pinaceae and Rubiaceae. Existing reviews of those families dealing with chemotaxonomy or natural product composition show that there are subfamilies, tribes or genera that are well studied, while there is no information for others.^{133–136} Although, those studies mainly focus on only a subset of natural product classes present in those families, they reflect the state of information that is accessible for those. However, the evolution theory suggests that related plants share related natural products with adaptations to their respective ecological niche. This might suggest that also the phylogenetic tree regions of a family which shows underrepresentation might produce similar natural products like the rest of the family.^{16,17,137} Consequently, it is of high probability to identify known natural products as well as derivatives, however, also new scaffolds are not excluded.

The fact that in most cases no absence data (negative results) are published for natural product studies, is a huge drawback for our study. There is no possibility to distinguish if there are regions in the phylogenetic tree that produce fewer natural products than

others or if they occur as underrepresented because they are only less studied than others. Nevertheless, this approach suggests some plant families with potential for yet unknown natural products.

5.2.4 Are there less studied regions?

Another question we addressed was if there are regions that are more or less studied than others and if it is really necessary to head to extreme condition habitats to bring the natural product research forward. Occurrences from the Global Biodiversity Information Facility (GBIF)⁷¹ were used to map the species identified in our literature approach onto a world map. To compare the spatial patterns that were found for the natural products from plants, they were compared to the spatial patterns that were obtained for all plants with occurrences from GBIF. Already the number of species based on all plant occurrences from GBIF (Figure 5.2A) shows, that it seems that there are regions, which are less studied, here in terms of plant occurrences. These are indeed regions of extreme environmental conditions like deserts in Northern Africa, or cold regions like Siberia, North Canada, Greenland or Antarctica, which however, are also not very species rich. The highest number of plants studied can be found in this analysis in Europe, Central America and some coastal regions and cannot be compared to other biodiversity illustrations which are based on species distribution models and not on sole occurrence data. However, the advantage of this representation is that it reflects the location of sample studies and reveals less studied regions.

The rasterized version of GBIF occurrences is also the basis of the analysis of plants for which we extracted natural product information from literature. The overall pattern (Figure 5.2B) is similar to all plant occurrences. An outstanding finding is that for each grid cell for which we had occurrence data of all plants in GBIF, we also identified plants in our text mining approach. Meaning that even we have probably only a minor part of available information collected from existing publications, this already may significantly reflect the distribution of studies worldwide. Of course, the total numbers of plants are much lower compared to all GBIF occurrences, therefore we studied the ratio of plants with natural product information to all GBIF plants per grid cell. This reveals regions where a high number of plants that have been reported there are also studied in terms of natural products like in South East Asia (see Figure A2). However, it reveals also regions where plant potential seems not to be used fully, like South Africa, Europe, Australia or the West coast of America. As a result one can conclude, that there are

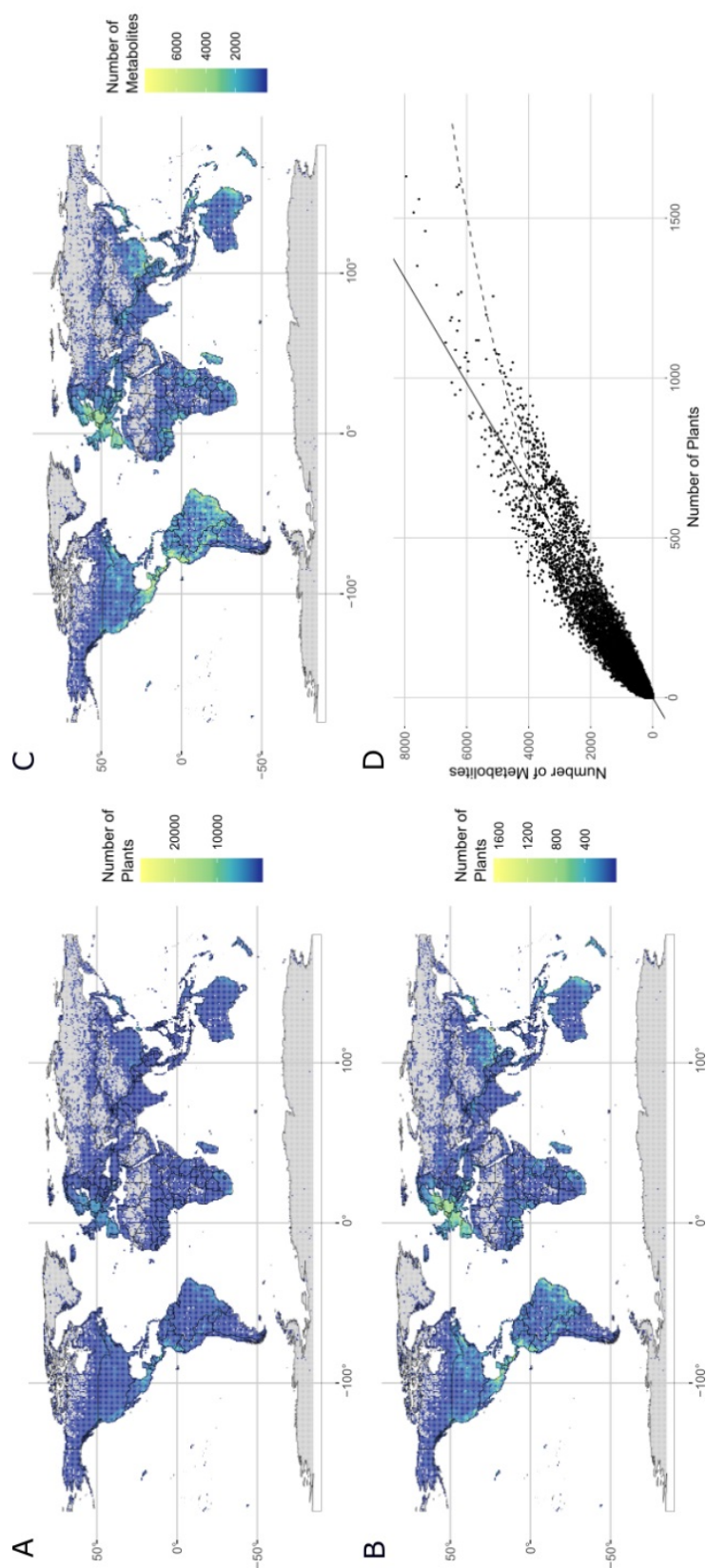


Figure 5.2: Bio- and chemodiversity of the world based on GBIF occurrences and natural product relations identified by text mining approach. The raster has a resolution of approximately 100 km per 100 km. For each grid cell the species or metabolites occurring there are counted separately. A) Number of unique plant (genus, species, subspecies) occurrences recorded in GBIF (accessed 1st April 2019). B) Number of species with metabolite information identified with text mining approach based on GBIF occurrences. C) Number of metabolites from the text mining approach. Here the occurrences of the plants from B were used to identify the grid cells where each metabolite occurs. D) Number of species (B) per grid cell versus the corresponding number of metabolites (C) together with a linear regression (solid line, $r^2=0.96$) and a fitted growth function (dashed line, $(M(p) = S - (S - M0)e^{-kp})$ with M being the number of metabolites and p being the number of plants).

indeed less studied region in terms of plant collection as well as in natural product research and it also shows that it is not necessary to head to extreme environmental conditions to investigate plants that are not yet studied in terms of natural products.

5.2.5 Are species-rich tropical rainforests really the source of most natural products?

Using the occurrences of plants already introduced above, one can also conclude in which habitats these plants grow, using habitat maps like the global terrestrial ecoregions map published by WWF⁷². Additionally, to the overall global pattern we were also interested if the number of plants per habitat correlate in terms of natural products studied. Figure 5.3 shows the number of plants for each habitat based on the GBIF occurrences and the ratio of the plants which have been identified in our text mining approach. Not surprisingly, most plants can be found in moist broadleaf forest, typical for tropical and subtropical lowlands and medium heights. However, for only 2.8 % of them we found publications on the chemical constituents. Inland water regions (21.6 %), flooded grasslands and savannas (15.4 %), rock and ice (11.6 %) and mangroves (10.1 %) are the habitats with the highest ratio of plants that were studied for natural products in the text mining approach. The remaining habitats show ratios of the text mined species to the GBIF species between 4.2 % and 8.3 %. We are aware of the fact that the percentages of the ratio may be higher if the collection of data was more comprehensive, however, due to our unbiased way to gain such data we are confident that the proportion values for the habitats among themselves are stable. Therefore, these ratios reflect already the finding of the study presented above, meaning that there are even plants and genera in easily accessible regions which are still not investigated in terms of natural products.

5.2.6 Do chemical and biological diversity go hand in hand?

Although biodiversity is not studied here in a classical way based on species distribution models, but based on pure occurrence data without modelling the distribution, the comparison of bio- and chemodiversity can give an insight if the chemical and biological diversity go hand in hand. Ramesha *et al.*⁸ introduce two different hypothesis how biodiversity and chemodiversity can be related. The first is a linear relation, where the chemical diversity constantly rises with increasing biodiversity, whereas the second suggests a limited growth curve, with a rapid increase of chemical diversity along with

biodiversity in the beginning and then the increase slows down to reach a maximum. A definitive answer which of the hypothesis describes the relation cannot be given, since only a minor part of the plants is investigated in terms of natural products.^{126,127} However, the studies of Pye *et al.*¹³⁸ suggest, that at least for the marine organisms there is still a relevant number of chemical compounds that are structurally distinct, but also that there is also a limited number of chemical scaffolds accessible from nature. The limitation of at least the scaffolds is in accordance to the evolutionary development of plants, which predetermines that metabolites will not be randomly distributed across species but that similar compounds will be found in closely related ones.^{14,17,18} Our own studies also do not answer the question in which way chemo- and biodiversity go

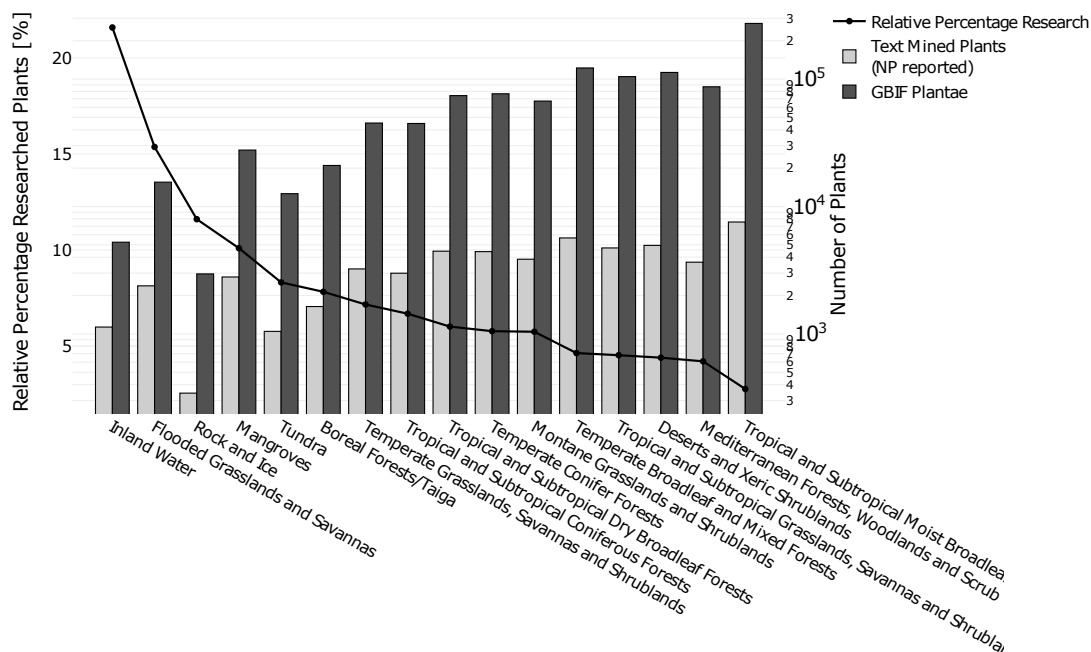


Figure 5.3: Distribution of plants per habitat. In dark grey the numbers of all plants with occurrence records in GBIF accessed on 1st April 2019 are shown and in light grey all plants (genus, species and subspecies) identified with the text mining approach along with their natural products. Please mind the logarithmic scale. The percentage gives the ratio of the text mined plants in comparison to all GBIF plants.

hand in hand, however they show some correlation (Figure 5.2D). But it is not possible to decide whether this is a purely linear correlation or still a linear increase of a curve that will go to a plateau of chemical diversity. Both possible functions were fitted to the data and both are significant ($p < 0.001$), but looking on the plot shows that the truth is somewhere inbetween. Although, one cannot say which function fits better it is worth to mention that there is a clear correlation of chemical and biological diversity in our data, which supports the above mentioned studies and shows again that our data reflect the current state of knowledge. The comparison of the maps of bio- and chemical diversity of our data (Figure 5.2B+C) emphasize this correlation. For the chemical diversity plot, all metabolites reported for all plants present per grid cell were counted uniquely.

5.3 Summary

Studying distribution and composition of natural products one should never forget that every plant produces natural products, but they differ in structure and amount. However, not every plant is studied for their natural products and here we present the recent state of knowledge about natural products in plants. The unbiased, even if not completely comprehensive, data collection guarantees new insights in natural product research. We were able to answer important questions and suggested ways of identifying answers for e.g. yet little studied regions as well as for plant families with potential value for future natural product research.

6 Discussion and Summary

After a successful proof of concept of the combination of spatial, phylogenetic and chemical data of plants from Indonesia, the second aim was an unbiased (no focus on plant families, regions or uses) collection of plants and their corresponding natural products. Therefore, a gold set was created, consisting of 102 full text publications. One drawback of this goldset is the query which was chosen to select the publications. In this query the term “Indonesia” was included. However, since the English terminology used to describe which natural products were extracted from a natural source is independent of the region where the source was collected or the researcher comes from, the gold set can be used to program a suitable tool for the relation extraction from publications. Another term used in the query was “plant”. Although, the terminology is independent in terms of the way of extraction, it is not in describing the natural source. Consequently, the gold set can be used to train named entity recognition of plants, but not of general natural sources (e.g. fungi, algae). Therefore, in order to collect more comprehensive data, the gold set should be extended to include publications with natural products from other organ systems. But also to include publications dealing with chemical plants or synthetic natural products to reduce the number of false positives. The relations not only annotated within sentences, but also spanning beyond sentences and within tables, are an advantage of this corpus in comparison to e.g. Choi *et al.*'s corpus⁴⁸. The developed relation extraction tool which was used to create the globally dataset is, beside pre-trained tagger (part of speech tagger from nltk⁹⁵ and the chemical tagger from the ChemDataExtractor⁹⁶), rule based. On the one hand the rule based approach is a big advantage since it is reasonable why a word is tagged as which entity or why the relation was extracted. On the other hand it is a disadvantage, because it is always a balance between the accuracy and the comprehensive of a rule. If rules are so accurate, that no false positives are selected, one would need an immense number of rules to cover most of relations. But the rules should also not be too general to have only a

hand full of rules to cover all relations because of the huge amount of false positives one would extract with this approach. Here machine learning for named entity recognition can help. There are ways to train models with the surrounding words which could be an improvement especially for the chemical named entity recognition. First tests were performed and showed indeed increased values for precision and recall. This could also be a way to enhance the relation extraction. Another possibility to improve the extraction would be to resort to artificial intelligence. Several groups have used the GPT¹⁴¹ models that ChatGPT⁵² uses or BART¹⁴³ models for information extraction tasks and have achieved promising results.¹⁴⁶ Of course, these also always need to be trained on a specific use case. For this purpose, the gold standard corpus presented here could be used to enable the extraction of natural products and their natural sources using AI. Since the parts of a plant and the region where the plant described in a publication was collected is not extracted with this text mining tool, and comparisons in which plant parts active compounds are produced as described by Chassagne *et al.*¹¹² are not possible. Also it is not (yet) possible to compare if one species grown in different habitats or regions produces different compounds. The spatial analysis shown in this thesis are based on general knowledge where the species grow (accessed via GBIF). Of course, the database created with a number of around 100,000 natural products is not the biggest one available. For example, the dictionary of natural product²¹ v27 consists of ~300,000 and the recently freely accessible published COCONUT database²⁹ of ~400,000 natural products. The developed tool and collected data should not be understood as stand-alone, it is meant as a useful tool to automatically keep databases up to date. Obviously, still manual work is required to create comprehensive databases, but the tool can reduce the amount of work. As mentioned above, the data used for the analysis in the last chapter might not be the most comprehensive ones, but it shows that the data that can be collected and seem to be unbiased concerning the phylogenetic and spatial information of plants and chemical classes of the natural products which was the main aim at the start in order to identify in an unbiased way regions of (chemical) interest or parts of the phylogenetic tree which are understudied compared to the average.

References

- (1) All Natural. *Nature Chemical Biology* **2007**, *3*, 351 DOI: 10.1038/nchembio0707-351.
- (2) Theis, N.; Lerdau, M. The Evolution of Function in Plant Secondary Metabolites. *International Journal of Plant Sciences* **2003**, *164* (S3), S93–S102.
- (3) Newman, D. J.; Cragg, G. M. Natural Products as Sources of New Drugs from 1981 to 2014. *Journal of Natural Products* **2016**, *79* (3), 629–661.
- (4) Fischbach, M. A.; Walsh, C. T. Antibiotics for Emerging Pathogens. *Science* **2009**, *325* (5944), 1089–1093 DOI: 10.1126/science.1176667.
- (5) Firn, R. D. Bioprospecting - Why Is It so Unrewarding? *Biodiversity and Conservation* **2003**, *12*, 207–216.
- (6) Katz, L.; Baltz, R. H. Natural Product Discovery: Past, Present, and Future. *Journal of Industrial Microbiology and Biotechnology* **2016**, *43* (2-3), 155–176.
- (7) Fabricant, D. S.; Farnsworth, N. R. The Value of Plants Used in Traditional Medicine for Drug Discovery. *Environmental Health Perspectives* **2001**, *109* (Suppl. 1), 69–75 DOI: 10.1289/ehp.01109s169.
- (8) Ramesha, B. T.; Gertsch, J.; Ravikanth, G.; Priti, V.; Ganeshaiyah, K. N.; Uma Shaanker, R. Biodiversity and Chemodiversity: Future Perspectives in Bioprospecting. *Current Drug Targets* **2011**, *12* (11), 1–16.
- (9) Kew, R. *The State of the World's Plants Report - 2016*; Report; Royal Botanic Gardens, Kew, 2016.
- (10) Verpoorte, R. Pharmacognosy in the New Millennium: Leadfinding and Biotechnology. *Journal of Pharmacy and Pharmacology* **2000**, *52* (3), 253–262 DOI: doi:10.1211/0022357001773931.
- (11) Mendelsohn, R.; Balick, M. J. The Value of Undiscovered Pharmaceuticals in Tropical Forests. *Economic Botany* **1995**, *49* (2), 223–228.
- (12) Berdy, J. Bioactive Microbial Metabolites. *The Journal of antibiotics* **2005**, *58* (1), 1–26.
- (13) Bérdy, J. Thoughts and Facts about Antibiotics: Where We Are Now and Where We Are Heading. *The Journal of antibiotics* **2012**, *65* (8), 385–395.

-
- (14) Hartmann, T. From Waste Products to Ecochemicals: Fifty Years Research of Plant Secondary Metabolism. *Phytochemistry* **2007**, *68*, 2831–2846 DOI: 10.1016/j.phytochem.2007.09.017.
- (15) Rønsted, N.; Chase, M. W.; Albach, D. C.; Bello, M. A. Phylogenetic Relationships Within *Plantago* (Plantaginaceae): Evidence from Nuclear Ribosomal ITS and Plastid trnL-f Sequence Data. *Botanical Journal of the Linnean Society* **2002**, *139* (4), 323–338.
- (16) Delgoda, R.; Murray, J. Evolutionary Perspectives on the Role of Plant Secondary Metabolites. In *Pharmacognosy*; Elsevier, 2017; pp 93–100.
- (17) Wink, M. Evolution of Secondary Metabolites from an Ecological and Molecular Phylogenetic Perspective. *Phytochemistry* **2003**, *64* (1), 3–19 DOI: [http://dx.doi.org/10.1016/S0031-9422\(03\)00300-5](http://dx.doi.org/10.1016/S0031-9422(03)00300-5).
- (18) Wink, M.; Mohamed, G. I. A. Evolution of Chemical Defense Traits in the Leguminosae: Mapping of Distribution Patterns of Secondary Metabolites on a Molecular Phylogeny Inferred from Nucleotide Sequences of the rbcL Gene. *Biochemical Systematics and Ecology* **2003**, *31* (8), 897–917 DOI: [http://dx.doi.org/10.1016/S0305-1978\(03\)00085-1](http://dx.doi.org/10.1016/S0305-1978(03)00085-1).
- (19) Zhu, F.; Qin, C.; Tao, L.; Liu, X.; Shi, Z.; Ma, X.; Jia, J.; Tan, Y.; Cui, C.; Lin, J. Clustered Patterns of Species Origins of Nature-Derived Drugs and Clues for Future Bioprospecting. *Proceedings of the National Academy of Sciences of the United States of America* **2011**, *108* (31), 12943–12948.
- (20) Bacteria, J. X.; Noge, K.; Venable, D. L. Macroevolutionary Chemical Escalation in an Ancient Plant–Herbivore Arms Race. *Proceedings of the National Academy of Sciences* **2009**, *106* (43), 18062–18066 DOI: 10.1073/pnas.0904456106.
- (21) Buckingham, J. Dictionary of Natural Products. **1993**.
- (22) Goodman, J. Computer Software Review: Reaxys, 2009.
- (23) Xue, R.; Fang, Z.; Zhang, M.; Yi, Z.; Wen, C.; Shi, T. TCMID: Traditional Chinese Medicine Integrative Database for Herb Molecular Mechanism Analysis. *Nucleic acids research* **2012**, *41* (D1), D1089–D1095.

- (24) Shinbo, Y.; Nakamura, Y.; Altaf-Ul-Amin, M.; Asahi, H.; Kurokawa, K.; Arita, M.; Saito, K.; Ohta, D.; Shibata, D.; Kanaya, S. KNApSAcK: A Comprehensive Species-Metabolite Relationship Database. In *Plant metabolomics*; Springer, 2006; pp 165–181.
- (25) Afendi, F. M.; Okada, T.; Yamazaki, M.; Hirai-Morita, A.; Nakamura, Y.; Nakamura, K.; Ikeda, S.; Takahashi, H.; Altaf-Ul-Amin, M.; Darusman, L. K. KNApSAcK Family Databases: Integrated Metabolite–Plant Species Databases for Multifaceted Plant Research. *Plant and Cell Physiology* **2012**, *53* (2), e1–e1.
- (26) Xie, T.; Song, S.; Li, S.; Ouyang, L.; Xia, L.; Huang, J. Review of Natural Product Databases. *Cell Prolif* **2015**, *48* (4), 398–404 DOI: 10.1111/cpr.12190.
- (27) Zeng, X.; Zhang, P.; He, W.; Qin, C.; Chen, S.; Tao, L.; Wang, Y.; Tan, Y.; Gao, D.; Wang, B.; Chen, Z.; Chen, W.; Jiang, Y. Y.; Chen, Y. Z. NPASS: Natural Product Activity and Species Source Database for Natural Product Research, Discovery and Tool Development. *Nucleic Acids Res* **2017** DOI: 10.1093/nar/gkx1026.
- (28) Zeng, Y. Z., X.; Chen. Natural Product Activity & Species Source Database (NPASS). [Http://Bidd2.nus.edu.sg/NPASS/](http://Bidd2.nus.edu.sg/NPASS/), 2017.
- (29) Sorokina, M.; Steinbeck, C. Review on Natural Products Databases: Where to Find Data in 2020. *Journal of cheminformatics* **2020**, *12* (1), 1–51.
- (30) Chen, J. Y.; Shen, C.; Sivachenko, A. Y. Mining Alzheimer Disease Relevant Proteins from Integrated Protein Interactome Data. In *Pac symp biocomput*; 2006.
- (31) Choi, W.; Lee, H. A Text Mining Approach for Identifying Herb-Chemical Relationships from Biomedical Articles. **2015**, 25–25 DOI: 10.1145/2811163.2811178.
- (32) Cohen, A. M.; Hersh, W. R. A Survey of Current Work in Biomedical Text Mining. *Briefings in bioinformatics* **2005**, *6* (1), 57–71.
- (33) Huang, C.-C.; Lu, Z. Community Challenges in Biomedical Text Mining over 10 Years: Success, Failure and the Future. *Briefings in bioinformatics* **2016**, *17* (1), 132–144.
- (34) Kanya, N.; Ravi, T. A Study on Relationship Extraction from Text Data. *Data Mining and Knowledge Engineering* **2016**, *8* (7), 216–221.

-
- (35) Kilicoglu, H. Biomedical Text Mining for Research Rigor and Integrity: Tasks, Challenges, Directions. *bioRxiv* **2017**, 108480.
- (36) Krallinger, M.; Rabal, O.; Lourenco, A.; Oyarzabal, J.; Valencia, A. Information Retrieval and Text Mining Technologies for Chemistry. *Chemical Reviews* **2017**.
- (37) Singhal, A.; Leaman, R.; Catlett, N.; Lemberger, T.; McEntyre, J.; Polson, S.; Xenarios, I.; Arighi, C.; Lu, Z. Pressing Needs of Biomedical Text Mining in Biocuration and Beyond: Opportunities and Challenges. *Database* **2016**, *2016*, baw161.
- (38) Singhal, A.; Simmons, M.; Lu, Z. Text Mining for Precision Medicine: Automating Disease-Mutation Relationship Extraction from Biomedical Literature. *Journal of the American Medical Informatics Association* **2016**, *23* (4), 766–772.
- (39) Vazquez, M.; Krallinger, M.; Leitner, F.; Valencia, A. Text Mining for Drugs and Chemical Compounds: Methods, Tools and Applications. *Molecular Informatics* **2011**, *30* (6-7), 506–519.
- (40) Blake, C. Beyond Genes, Proteins, and Abstracts: Identifying Scientific Claims from Full-Text Biomedical Articles. *J Biomed Inform* **2010**, *43* (2), 173–189 DOI: 10.1016/j.jbi.2009.11.001.
- (41) Westergaard, D.; Staerfeldt, H. H.; Tonsberg, C.; Jensen, L. J.; Brunak, S. A Comprehensive and Quantitative Comparison of Text-Mining in 15 Million Full-Text Articles Versus Their Corresponding Abstracts. *PLoS Comput Biol* **2018**, *14* (2), e1005962 DOI: 10.1371/journal.pcbi.1005962.
- (42) Krallinger, M.; Rabal, O.; Leitner, F.; Vazquez, M.; Salgado, D.; Lu, Z.; Leaman, R.; Lu, Y.; Ji, D.; Lowe, D. M. The CHEMDNER Corpus of Chemicals and Drugs and Its Annotation Principles. *Journal of cheminformatics* **2015**, *7* (1), S2.
- (43) Kim, J.-D.; Ohta, T.; Tateisi, Y.; Tsujii, J. GENIA Corpus—a Semantically Annotated Corpus for Bio-Textmining. *Bioinformatics* **2003**, *19* (suppl 1), i180–i182.
- (44) Pyysalo, S.; Ginter, F.; Heimonen, J.; Bjorne, J.; Boberg, J.; Jarvinen, J.; Salakoski, T. BioInfer: A Corpus for Information Extraction in the Biomedical Domain. *BMC Bioinformatics* **2007**, *8*, 50 DOI: 10.1186/1471-2105-8-50.

- (45) Islamaj Dogan, R.; Kim, S.; Chatr-Aryamontri, A.; Chang, C. S.; Oughtred, R.; Rust, J.; Wilbur, W. J.; Comeau, D. C.; Dolinski, K.; Tyers, M. The BioC-BioGRID Corpus: Full Text Articles Annotated for Curation of Protein-Protein and Genetic Interactions. *Database (Oxford)* **2017**, 2017 DOI: 10.1093/database/baw147.
- (46) Jensen, K.; Panagiotou, G.; Kouskoumvekaki, I. Integrated Text Mining and Chemoinformatics Analysis Associates Diet to Health Benefit at Molecular Level. *PLoS Comput Biol* **2014**, 10 (1), e1003432.
- (47) Choi, W.; Choi, C. H.; Kim, Y. R.; Kim, S. J.; Na, C. S.; Lee, H. HerDing: Herb Recommendation System to Treat Diseases Using Genes and Chemicals. *Database (Oxford)* **2016**, 2016 DOI: 10.1093/database/baw011.
- (48) Choi, W.; Kim, B.; Cho, H.; Lee, D.; Lee, H. A Corpus for Plant-Chemical Relationships in the Biomedical Domain. *BMC Bioinformatics* **2016**, 17, 386 DOI: 10.1186/s12859-016-1249-5.
- (49) Shardlow, M.; Nguyen, N.; Owen, G.; O'Donovan, C.; Leach, A.; McNaught, J.; Turner, S.; Ananiadou, S. A New Corpus to Support Text Mining for the Curation of Metabolites in the ChEBI Database. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*; pp 280–285.
- (50) Ellis, E. C.; Antill, E. C.; Kreft, H. All Is Not Loss: Plant Biodiversity in the Anthropocene. *PloS one* **2012**, 7 (1).
- (51) Kahle, D.; Wickham, H. Ggmap: Spatial Visualization with Ggplot2. *The R Journal* **2013**, 5 (1), 144–161.
- (52) OpenAI. ChatGPT.
- (53) Mittermeier, R. A.; Mittermeier, C. G. *Megadiversity: Earth's Biologically Wealthiest Nations*; Cemex, 1997.
- (54) Holzmeyer, L.; Hartig, A.-K.; Franke, K.; Brandt, W.; Muellner-Riehl, A. N.; Wessjohann, L. A.; Schnitzler, J. Evaluation of Plant Sources for Antiinfective Lead Compound Discovery by Correlating Phylogenetic, Spatial, and Bioactivity Data. *Proceedings of the National Academy of Sciences* **2020**, 117 (22), 12444–12451.

-
- (55) Backer, C. A.; Bakhuizen van den Brink, R. C. *Flora of Java*; The Rijksherbarium: Leyden, NL, 1963-1968.
- (56) Priyadi, H.; Takao, G.; Rahmawati, I.; Supriyanto, B.; Nursal, W. I.; Rahman, I. *Five Hundred Plant Species in Gunung Halimun Salak National Park, West Java: A Checklist Including Sundanese Names, Distribution, and Use*; CIFOR, 2010.
- (57) Boyle, B.; Hopkins, N.; Lu, Z.; Raygoza Garay, J. A.; Mozzherin, D.; Rees, T.; Matasci, N.; Narro, M. L.; Piel, W. H.; McKay, S. J.; Lowry, S.; Freeland, C.; Peet, R. K.; Enquist, B. J. The Taxonomic Name Resolution Service: An Online Tool for Automated Standardization of Plant Names. *BMC Bioinformatics* **2013**, *14*, 16 DOI: 10.1186/1471-2105-14-16.
- (58) TNRS (2018) iPlant Collaborative.
- (59) Reitz, K. Python Package Requests 2.10.0, 2016.
- (60) Richardson, L. Python Package BeautifulSoup4 4.5.1, 2016.
- (61) Lee, J. J. Python Package Mechanize 0.2.5, 2011.
- (62) Muthukadan, B. Python Package Selenium 3.8.1, 2018.
- (63) ULC, C. C. G. Molecular Operating Environment (MOE), 2018.
- (64) Kanehisa, M.; Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **2000**, *28* (1), 27–30 DOI: 10.1093/nar/28.1.27.
- (65) Morishima, K.; Tanabe, M.; Furumichi, M.; Kanehisa, M.; Sato, Y. KEGG: New Perspectives on Genomes, Pathways, Diseases and Drugs. *Nucleic Acids Research* **2016**, *45* (D1), D353–D361 DOI: 10.1093/nar/gkw1092.
- (66) Morishima, K.; Tanabe, M.; Furumichi, M.; Kanehisa, M.; Sato, Y. New Approach for Understanding Genome Variations in KEGG. *Nucleic Acids Research* **2018**, *47* (D1), D590–D595 DOI: 10.1093/nar/gky962.
- (67) Reitz, K. Python Package Requests 2.19.1, 2018.

- (68) Contributors, T. B. Python Package Bio 1.71, 2018.
- (69) Hinchliff, C. E.; Smith, S. A. Some Limitations of Public Sequence Data for Phylogenetic Inference (in Plants). *PLoS ONE* **2014**, *9* (7), e98986 DOI: 10.1371/journal.pone.0098986.
- (70) Benson, D. A.; Karsch-Mizrachi, I.; Lipman, D. J.; Ostell, J.; Wheeler, D. L. GenBank. *Nucleic acids research* **2008**, *36*, D25.
- (71) GBIF.org. GBIF Occurrence Download (Accessed: 1st April 2019) <https://doi.org/10.15468/Dl.plma38>.
- (72) Olson, D. M.; Dinerstein, E.; Wikramanayake, E. D.; Burgess, N. D.; Powell, G. V.; Underwood, E. C.; D'Amico, J. A.; Itoua, I.; Strand, H. E.; Morrison, J. C. Terrestrial Ecoregions of the World: A New Map of Life on Earth a New Global Map of Terrestrial Ecoregions Provides an Innovative Tool for Conserving Biodiversity. *BioScience* **2001**, *51* (11), 933–938.
- (73) Ciccarelli, F. D.; Doerks, T.; Von Mering, C.; Creevey, C. J.; Snel, B.; Bork, P. Toward Automatic Reconstruction of a Highly Resolved Tree of Life. *science* **2006**, *311* (5765), 1283–1287.
- (74) Letunic, I.; Bork, P. Interactive Tree of Life (iTOL) V3: An Online Tool for the Display and Annotation of Phylogenetic and Other Trees. *Nucleic acids research* **2016**, *44* (W1), W242–W245.
- (75) Fritz, S. A.; Purvis, A. Selectivity in Mammalian Extinction Risk and Threat Types: A New Measure of Phylogenetic Signal Strength in Binary Traits. *Conservation Biology* **2010**, *24* (4), 1042–1051 DOI: 10.1111/j.1523-1739.2010.01455.x.
- (76) Pearse, D. O. R. F. G. T. T. P. S. F. N. I. W. R Package Caper 0.5.2, 2013.
- (77) Losos, J. B. Phylogenetic Niche Conservatism, Phylogenetic Signal and the Relationship Between Phylogenetic Relatedness and Ecological Similarity Among Species. *Ecology Letters* **2008**, *11* (10), 995–1003.
- (78) Webb, C. O.; Ackerly, D. D.; Kembel, S. W. Phylocom: Software for the Analysis of Phylogenetic Community Structure and Trait Evolution. *Bioinformatics* **2008**, *24* (18), 2098–2100 DOI: btn358 [pii] 10.1093/bioinformatics/btn358.
- (79) GBIF.org. GBIF Occurrence Download (Accessed: 3rd May 2018) <https://doi.org/10.15468/Dl.jj7fsg>.

- (80) Gratton, P.; Marta, S.; Bocksberger, G.; Winter, M.; Trucchi, E.; Köhl, H. A World of Sequences: Can We Use Georeferenced Nucleotide Databases for a Robust Automated Phylogeography? *Journal of Biogeography* **2017**, *44* (2), 475–486 DOI: doi:10.1111/jbi.12786.
- (81) Hernandez, P. A.; Graham, C. H.; Master, L. L.; Albert, D. L. The Effect of Sample Size and Species Characteristics on Performance of Different Species Distribution Modeling Methods. *Ecography* **2006**, *29* (5), 773–785.
- (82) Proosdij, A. S. van; Sosef, M. S.; Wieringa, J. J.; Raes, N. Minimum Required Number of Specimen Records to Develop Accurate Species Distribution Models. *Ecography* **2016**, *39* (6), 542–552.
- (83) Schmitt, S.; Pouteau, R.; Justeau, D.; Boissieu, F.; Birnbaum, P. Ssdm: An R Package to Predict Distribution of Species Richness and Composition Based on Stacked Species Distribution Models. *Methods in Ecology and Evolution* **2017**, *8* (12), 1795–1803 DOI: doi:10.1111/2041-210X.12841.
- (84) Barbet-Massin, M.; Jiguet, F.; Albert, C. H.; Thuiller, W. Selecting Pseudo-absences for Species Distribution Models: How, Where and How Many? *Methods in ecology and evolution* **2012**, *3* (2), 327–338.
- (85) Allouche, O.; Tsoar, A.; Kadmon, R. Assessing the Accuracy of Species Distribution Models: Prevalence, Kappa and the True Skill Statistic (TSS). *Journal of Applied Ecology* **2006**, *43* (6), 1223–1232.
- (86) Hijmans, R. J.; Cameron, S. E.; Parra, J. L.; Jones, P. G.; Jarvis, A. Very High Resolution Interpolated Climate Surfaces for Global Land Areas. *International Journal of Climatology* **2005**, *25*, 1965–1978.
- (87) Hengl, T.; Mendes de Jesus, J.; Heuvelink, G. B. M.; Ruiperez Gonzalez, M.; Kilibarda, M.; Blagotić, A.; Shangquan, W.; Wright, M. N.; Geng, X.; Bauer-Marschallinger, B.; Guevara, M. A.; Vargas, R.; MacMillan, R. A.; Batjes, N. H.; Leenaars, J. G. B.; Ribeiro, E.; Wheeler, I.; Mantel, S.; Kempen, B. SoilGrids250m: Global Gridded Soil Information Based on Machine Learning. *PLOS ONE* **2017**, *12* (2), e0169748 DOI: 10.1371/journal.pone.0169748.

- (88) Pottier, J.; Dubuis, A.; Pellissier, L.; Maiorano, L.; Rossier, L.; Randin, C. F.; Vittoz, P.; Guisan, A. The Accuracy of Plant Assemblage Prediction from Species Distribution Models Varies Along Environmental Gradients. *Global Ecology and Biogeography* **2013**, *22* (1), 52–63.
- (89) Rønsted, N.; Symonds, M. R. E.; Birkholm, T.; Christensen, S. B.; Meerow, A. W.; Molander, M.; Mølgaard, P.; Petersen, G.; Rasmussen, N.; Staden, J. van; Stafford, G. I.; Jäger, A. K. Can Phylogeny Predict Chemical Diversity and Potential Medicinal Activity of Plants? A Case Study of Amaryllidaceae. *BMC Evolutionary Biology* **2012**, *12* (1), 182 DOI: 10.1186/1471-2148-12-182.
- (90) Böhme, T.; Irmer, M.; Püschel, A.; Bobach, C.; Laube, U.; Weber, L. OCMiner: Text Processing, Annotation and Relation Extraction for the Life Sciences. In *SWAT4LS*; Citeseer.
- (91) Roberts, R. J. PubMed Central: The GenBank of the Published Literature, 2001.
- (92) Stenetorp, P.; Pyysalo, S.; Topić, G.; Ohta, T.; Ananiadou, S.; Tsujii, J. BRAT: A Web-Based Tool for NLP-Assisted Text Annotation. In *Proceedings of the demonstrations at the 13th conference of the european chapter of the association for computational linguistics*; Association for Computational Linguistics; pp 102–107.
- (93) Schlaf, A.; Bobach, C.; Irmer, M. Creating a Gold Standard Corpus for the Extraction of Chemistry-Disease Relations from Patent Texts. In *LREC*; pp 2057–2061.
- (94) Kalwij, J. M. Review of “the Plant List, a Working List of All Plant Species.” *Journal of Vegetation Science* **2012**, *23* (5), 998–1002.
- (95) Loper, E.; Bird, S. Nltk: The Natural Language Toolkit. *arXiv preprint cs/0205028* **2002**.
- (96) Swain, M. C.; Cole, J. M. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *Journal of chemical information and modeling* **2016**, *56* (10), 1894–1904.
- (97) (2017), G. S. GBIF Backbone Taxonomy. Checklist Dataset <https://doi.org/10.15468/39omei> Accessed via GBIF.org on 2017-10-24.
- (98) Mozzherin, D., Dmitry; Shorthouse. Global Names Resolver. Copyright (c) 2012-2013 Marine Biological Laboratory., 2013.

- (99) Chamberlain, S. A.; Szöcs, E. Taxize: Taxonomic Search and Retrieval in r. *F1000Research* **2013**, *2*.
- (100) Feunang, Y. D.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; others. ClassyFire: Automated Chemical Classification with a Comprehensive, Computable Taxonomy. *Journal of cheminformatics* **2016**, *8* (1), 1–20.
- (101) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; others. PubChem Substance and Compound Databases. *Nucleic acids research* **2016**, *44* (D1), D1202–D1213.
- (102) Wohlgemuth, G.; Haldiya, P. K.; Willighagen, E.; Kind, T.; Fiehn, O. The Chemical Translation Service—a Web-Based Tool to Improve Standardization of Metabolomic Reports. *Bioinformatics* **2010**, *26* (20), 2647–2648.
- (103) Swain, M.; Group, N. NCI/CADD Chemical Identifier Resolver, 2009.
- (104) Sitzmann, M. NCI/CADD Chemical Identifier Resolver, 2016.
- (105) Lowe, D. M.; Corbett, P. T.; Murray-Rust, P.; Glen, R. C. Chemical Name to Structure: OPSIN, an Open Source Solution, 2011.
- (106) O’Boyle, N. M.; Hutchison, G. R. Cinfony—Combining Open Source Cheminformatics Toolkits Behind a Common Interface. *Chemistry Central Journal* **2008**, *2* (1), 1–10.
- (107) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; others. The ChEMBL Database in 2017. *Nucleic acids research* **2017**, *45* (D1), D945–D954.
- (108) EMBL-EBI. ChEMBL Database Version 24.1: [Doi.org/10.6019/CHEMBL.database.24.1](https://doi.org/10.6019/CHEMBL.database.24.1).
- (109) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.
- (110) Chollet, F.; others. Keras <https://github.com/fchollet/keras>.

- (111) Landrum, G. Rdkit Documentation. *Release* **2013**, 1 (1-79), 4.
- (112) Chassagne, F.; Samarakoon, T.; Porras, G.; Lyles, J. T.; Dettweiler, M.; Marquez, L.; Salam, A. M.; Shabih, S.; Farrokhi, D. R.; Quave, C. L. A Systematic Review of Plants with Antibacterial Activities: A Taxonomic and Phylogenetic Perspective. *Frontiers in pharmacology* **2021**, 11, 2069.
- (113) Chassagne, F.; Cabanac, G.; Hubert, G.; David, B.; Marti, G. The Landscape of Natural Product Diversity and Their Pharmacological Relevance from a Focus on the Dictionary of Natural Products®. *Phytochemistry Reviews* **2019**, 18 (3), 601–622.
- (114) Wetzels, S.; Schuffenhauer, A.; Roggo, S.; Ertl, P.; Waldmann, H. Cheminformatic Analysis of Natural Products and Their Chemical Space. *Chimia* **2007**, 61 (6), 355–355.
- (115) Rosén, J.; Gottfries, J.; Muresan, S.; Backlund, A.; Oprea, T. I. Novel Chemical Space Exploration via Natural Products. *Journal of medicinal chemistry* **2009**, 52 (7), 1953–1962.
- (116) Gurulingappa, H.; Mudi, A.; Toldo, L.; Hofmann-Apitius, M.; Bhate, J. Challenges in Mining the Literature for Chemical Information. *Rsc Advances* **2013**, 3 (37), 16194–16211.
- (117) Filippov, I. V.; Nicklaus, M. C. Optical Structure Recognition Software to Recover Chemical Information: OSRA, an Open Source Solution, 2009.
- (118) Rakotondraibe, L. H.; Graupner, P. R.; Xiong, Q.; Olson, M.; Wiley, J. D.; Krai, P.; Brodie, P. J.; Callmander, M. W.; Rakotobe, E.; Ratovoson, F.; others. Neolignans and Other Metabolites from *Ocotea cymosa* from the Madagascar Rain Forest and Their Biological Activities. *Journal of Natural Products* **2015**, 78 (3), 431–440.
- (119) Honnibal, M.; Montani, I. Python Package Spacy 2.0, 2018.
- (120) Roberts, R. J. PubMed Central: The GenBank of the Published Literature, 2001.
- (121) MEDLINE, 1879-present.

- (122) Hastings, J.; Matos, P. de; Dekker, A.; Ennis, M.; Harsha, B.; Kale, N.; Muthukrishnan, V.; Owen, G.; Turner, S.; Williams, M. The ChEBI Reference Database and Ontology for Biologically Relevant Chemistry: Enhancements for 2013. *Nucleic acids research* **2012**, *41* (D1), D456–D463.
- (123) Rogers, F. B. Medical Subject Headings. *Bulletin of the Medical Library Association* **1963**, *51* (1), 114–116.
- (124) McNeill, J.; Barrie, F.; Buck, W.; Demoulin, V.; Greuter, W.; Hawksworth, D.; Herendeen, P.; Knapp, S.; Marhold, K.; Prado, J.; others. *International Code of Nomenclature for Algae, Fungi and Plants (Melbourne Code)*; Koeltz Scientific Books Königstein, 2012; Vol. 154.
- (125) Hinchliff, C. E.; Smith, S. A. Some Limitations of Public Sequence Data for Phylogenetic Inference (in Plants). *PLoS ONE* **2014**, *9* (7), e98986 DOI: 10.1371/journal.pone.0098986.
- (126) Dias, D. A.; Urban, S.; Roessner, U. A Historical Overview of Natural Products in Drug Discovery. *Metabolites* **2012**, *2* (2), 303–336.
- (127) Cragg, G. M.; Newman, D. J. Biodiversity: A Continuing Source of Novel Drug Leads. *Pure and applied chemistry* **2005**, *77* (1), 7–24.
- (128) Webb, C. O.; Ackerly, D. D.; Kembel, S. W. Phylocom: Software for the Analysis of Phylogenetic Community Structure and Trait Evolution. *Bioinformatics* **2008**, *24* (18), 2098–2100 DOI: btn358 [pii] 10.1093/bioinformatics/btn358.
- (129) Xie, C.-F.; Lou, H.-X. Secondary Metabolites in Bryophytes: An Ecological Aspect. *Chemistry & biodiversity* **2009**, *6* (3), 303–312.
- (130) Cove, D.; Bezanilla, M.; Harries, P.; Quatrano, R. Mosses as Model Systems for the Study of Metabolism and Development. *Annu. Rev. Plant Biol.* **2006**, *57*, 497–520.
- (131) Klavina, L.; Springe, G.; Nikolajeva, V.; Martsinkevich, I.; Nakurte, I.; Dzabijeva, D.; Steinberga, I. Chemical Composition Analysis, Antimicrobial Activity and Cytotoxicity Screening of Moss Extracts (Moss Phytochemistry). *Molecules* **2015**, *20* (9), 17221–17243.
- (132) Chen, F.; Ludwiczuk, A.; Wei, G.; Chen, X.; Crandall-Stotler, B.; Bowman, J. L. Terpenoid Secondary Metabolites in Bryophytes: Chemical Diversity, Biosynthesis and Biological Functions. *Critical Reviews in Plant Sciences* **2018**, *37* (2-3), 210–231.

- (133) Wink, M. Evolution of Secondary Metabolites in Legumes (Fabaceae). *South African Journal of Botany* **2013**, *89*, 164–175.
- (134) Martins, D.; Nunez, C. V. Secondary Metabolites from Rubiaceae Species. *Molecules* **2015**, *20* (7), 13422–13495.
- (135) Alvarenga, S. A. V.; Gastmans, J. P.; Vale Rodrigues, G. do; Moreno, P. R. H.; Paulo Emerenciano, V. de. A Computer-Assisted Approach for Chemotaxonomic Studies—Diterpenes in Lamiaceae. *Phytochemistry* **2001**, *56* (6), 583–595.
- (136) Wolff, R. L.; Deluc, L. G.; Marpeau, A. M.; Comps, B. Chemotaxonomic Differentiation of Conifer Families and Genera Based on the Seed Oil Fatty Acid Compositions: Multivariate Analyses. *Trees* **1997**, *12* (2), 57–65.
- (137) Weng, J.-K.; Philippe, R. N.; Noel, J. P. The Rise of Chemodiversity in Plants. *Science* **2012**, *336* (6089), 1667–1670 DOI: 10.1126/science.1217411.
- (138) Pye, C. R.; Bertin, M. J.; Lokey, R. S.; Gerwick, W. H.; Linington, R. G. Retrospective Analysis of Natural Products Provides Insights for Future Discovery Trends. *Proceedings of the National Academy of Sciences* **2017**, *114* (22), 5601–5606.
- (139) Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; others. Language Models Are Few-Shot Learners. *Advances in neural information processing systems* **2020**, *33*, 1877–1901.
- (140) Artetxe, M.; Bhosale, S.; Goyal, N.; Mihaylov, T.; Ott, M.; Shleifer, S.; Lin, X. V.; Du, J.; Iyer, S.; Pasunuru, R.; others. Efficient Large Scale Language Modeling with Mixtures of Experts. *arXiv preprint arXiv:2112.10684* **2021**.
- (141) Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; others. Training Language Models to Follow Instructions with Human Feedback. *Advances in Neural Information Processing Systems* **2022**, *35*, 27730–27744.
- (142) Paolini, G.; Athiwaratkun, B.; Krone, J.; Ma, J.; Achille, A.; Anubhai, R.; Santos, C. N. dos; Xiang, B.; Soatto, S. Structured Prediction as Translation Between Augmented Natural Languages. *arXiv preprint arXiv:2101.05779* **2021**.

-
- (143) Cabot, P.-L. H.; Navigli, R. REBEL: Relation Extraction by End-to-End Language Generation. In *Findings of the association for computational linguistics: EMNLP 2021*; 2021; pp 2370–2381.
- (144) Wan, Z.; Cheng, F.; Mao, Z.; Liu, Q.; Song, H.; Li, J.; Kurohashi, S. Gpt-Re: In-Context Learning for Relation Extraction Using Large Language Models. *arXiv preprint arXiv:2305.02105* **2023**.
- (145) Wadhwa, S.; Amir, S.; Wallace, B. C. Revisiting Relation Extraction in the Era of Large Language Models. *arXiv preprint arXiv:2305.05003* **2023**.
- (146) Li, B.; Fang, G.; Yang, Y.; Wang, Q.; Ye, W.; Zhao, W.; Zhang, S. Evaluating ChatGPT’s Information Extraction Capabilities: An Assessment of Performance, Explainability, Calibration, and Faithfulness. *arXiv preprint arXiv:2304.11633* **2023**.

Appendix

Table A1: KEGG metabolic pathways included in the selection of secondary metabolites

KEGG Identifier	Metabolic Pathway
1200	Carbon metabolism
1210	2-Oxocarboxylic acid metabolism
1212	Fatty acid metabolism
1230	Biosynthesis of amino acids
10	Glycolysis / Gluconeogenesis
20	Citrate cycle (TCA cycle)
30	Pentose phosphate pathway
40	Pentose and glucuronate interconversions
51	Fructose and mannose metabolism
52	Galactose metabolism
53	Ascorbate and aldarate metabolism
500	Starch and sucrose metabolism
520	Amino sugar and nucleotide sugar metabolism
620	Pyruvate metabolism
630	Glyoxylate and dicarboxylate metabolism
640	Propanoate metabolism
650	Butanoate metabolism
660	C5-Branched dibasic acid metabolism
562	Inositol phosphate metabolism
190	Oxidative phosphorylation
195	Photosynthesis
196	Photosynthesis - antenna proteins
710	Carbon fixation in photosynthetic organisms
720	Carbon fixation pathways in prokaryotes

KEGG Identifier	Metabolic Pathway
680	Methane metabolism
910	Nitrogen metabolism
920	Sulfur metabolism
61	Fatty acid biosynthesis
62	Fatty acid elongation
71	Fatty acid degradation
72	Synthesis and degradation of ketone bodies
73	Cutin, suberine and wax biosynthesis
100	Steroid biosynthesis
120	Primary bile acid biosynthesis
121	Secondary bile acid biosynthesis
140	Steroid hormone biosynthesis
561	Glycerolipid metabolism
564	Glycerophospholipid metabolism
565	Ether lipid metabolism
600	Sphingolipid metabolism
590	Arachidonic acid metabolism
591	Linoleic acid metabolism
592	alpha-Linolenic acid metabolism
1040	Biosynthesis of unsaturated fatty acids
230	Purine metabolism
240	Pyrimidine metabolism
250	Alanine, aspartate and glutamate metabolism
260	Glycine, serine and threonine metabolism
270	Cysteine and methionine metabolism
280	Valine, leucine and isoleucine degradation
290	Valine, leucine and isoleucine biosynthesis
300	Lysine biosynthesis

KEGG Identifier	Metabolic Pathway
310	Lysine degradation
220	Arginine biosynthesis
330	Arginine and proline metabolism
340	Histidine metabolism
350	Tyrosine metabolism
360	Phenylalanine metabolism
380	Tryptophan metabolism
400	Phenylalanine, tyrosine and tryptophan biosynthesis
410	beta-Alanine metabolism
430	Taurine and hypotaurine metabolism
440	Phosphonate and phosphinate metabolism
450	Selenocompound metabolism
460	Cyanoamino acid metabolism
471	D-Glutamine and D-glutamate metabolism
472	D-Arginine and D-ornithine metabolism
473	D-Alanine metabolism
480	Glutathione metabolism
510	N-Glycan biosynthesis
513	Various types of N-glycan biosynthesis
512	Mucin type O-glycan biosynthesis
515	Mannose type O-glycan biosynthesis
514	Other types of O-glycan biosynthesis
532	Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate
534	Glycosaminoglycan biosynthesis - heparan sulfate / heparin
533	Glycosaminoglycan biosynthesis - keratan sulfate
531	Glycosaminoglycan degradation

KEGG Identifier	Metabolic Pathway
563	Glycosylphosphatidylinositol (GPI)-anchor biosynthesis
601	Glycosphingolipid biosynthesis - lacto and neolacto series
603	Glycosphingolipid biosynthesis - globo and isoglobo series
604	Glycosphingolipid biosynthesis - ganglio series
540	Lipopolysaccharide biosynthesis
550	Peptidoglycan biosynthesis
511	Other glycan degradation
730	Thiamine metabolism
740	Riboflavin metabolism
750	Vitamin B6 metabolism
760	Nicotinate and nicotinamide metabolism
770	Pantothenate and CoA biosynthesis
780	Biotin metabolism
785	Lipoic acid metabolism
790	Folate biosynthesis
670	One carbon pool by folate
830	Retinol metabolism
860	Porphyrin and chlorophyll metabolism
130	Ubiquinone and other terpenoid-quinone biosynthesis
905	Brassinosteroid biosynthesis
981	Insect hormone biosynthesis
1070	Biosynthesis of plant hormones
3020	RNA polymerase
3022	Basal transcription factors
3040	Spliceosome
3010	Ribosome
970	Aminoacyl-tRNA biosynthesis

KEGG Identifier	Metabolic Pathway
3013	RNA transport
3015	mRNA surveillance pathway
3008	Ribosome biogenesis in eukaryotes
3060	Protein export
4141	Protein processing in endoplasmic reticulum
4130	SNARE interactions in vesicular transport
4120	Ubiquitin mediated proteolysis
4122	Sulfur relay system
3050	Proteasome
3018	RNA degradation
3030	DNA replication
3410	Base excision repair
3420	Nucleotide excision repair
3430	Mismatch repair
3440	Homologous recombination
3450	Non-homologous end-joining
3460	Fanconi anemia pathway
4144	Endocytosis
4145	Phagosome
4142	Lysosome
4146	Peroxisome
4140	Autophagy - animal
4138	Autophagy - yeast
4136	Autophagy - other
4137	Mitophagy - animal
4139	Mitophagy - yeast
4110	Cell cycle
4111	Cell cycle - yeast
4112	Cell cycle - Caulobacter
4113	Meiosis - yeast
4114	Oocyte meiosis

KEGG Identifier	Metabolic Pathway
4210	Apoptosis
4214	Apoptosis - fly
4215	Apoptosis - multiple species
4216	Ferroptosis
4217	Necroptosis
4115	p53 signaling pathway
4218	Cellular senescence
4640	Hematopoietic cell lineage
4610	Complement and coagulation cascades
4611	Platelet activation
4620	Toll-like receptor signaling pathway
4624	Toll and Imd signaling pathway
4621	NOD-like receptor signaling pathway
4622	RIG-I-like receptor signaling pathway
4623	Cytosolic DNA-sensing pathway
4625	C-type lectin receptor signaling pathway
4650	Natural killer cell mediated cytotoxicity
4612	Antigen processing and presentation
4660	T cell receptor signaling pathway
4658	Th1 and Th2 cell differentiation
4659	Th17 cell differentiation
4657	IL-17 signaling pathway
4662	B cell receptor signaling pathway
4664	Fc epsilon RI signaling pathway
4666	Fc gamma R-mediated phagocytosis
4670	Leukocyte transendothelial migration
4672	Intestinal immune network for IgA production
4062	Chemokine signaling pathway

Table A2: SMARTS strings, additional requirements and visualization of SMARTS as MOE descriptors for different natural product classes.



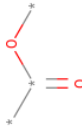
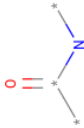


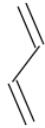
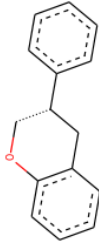
Natural product class	SMARTS (Additional requirements)	Visualization of SMARTS
terpene	<chem>[c,C]C(C)(C)C</chem> (a multiple of five connected carbon atoms is necessary, at least 10 carbons)	
benzofuran	<chem>[O,o]1c2c(cccc2)[#6][#6]1</chem>	
lacton	<chem>[r#6][r#6](O[r#6])=O</chem>	
lactam	<chem>[r#6][r#6](N[r#6])=O</chem>	
polyketide	<chem>[#6][#6]([#6][#6]=O)=O</chem>	
alkyl chain	<chem>[!R;C][!R;C]</chem> (compound is not classified as another natural product class)	
polyacetylene	<chem>[#6]=[#6][#6]=[#6]</chem>	
isoflavonoid	<chem>c1(C2C(Oc3c(cccc3)C2)cccc1</chem>	

Table A2: continued

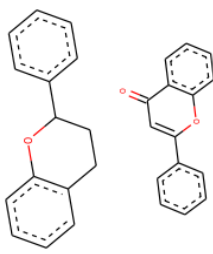
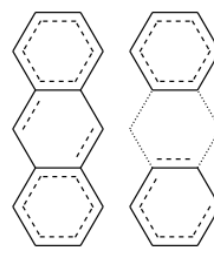
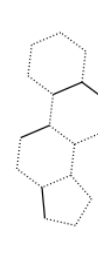
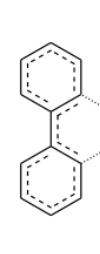

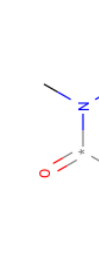
Natural product class	SMARTS (Additional requirements)	Visualization of SMARTS
flavonoids	<chem>c1(C2Oc3c(cccc3)CC2)cccc1</chem> <chem>O=C1c2c(OC(c3cccc3)=C1)cccc2</chem>	
anthracene	<chem>c12c(cc3c(c1)cccc3)cccc2</chem> <chem>c1(ccccc1[#6]2)[#6]c3c2cccc3</chem>	
steroids	<chem>C12C̃C̃C̃C̃1C̃3C̃(C̃(C̃C̃C̃4)C̃4C̃C̃3)C̃C̃2</chem>	
phenanthrene	<chem>c12c([#6][#6]c3cccc32)cccc1</chem>	
benzene	<chem>c1cccc1</chem> (compound is not classified as another natural product class)	
peptoides	<chem>C[!r:5;C](N(!:#1)C)=O</chem>	

Table A2: continued

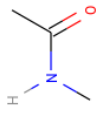
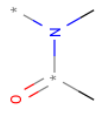


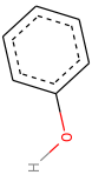
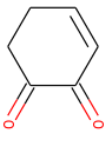
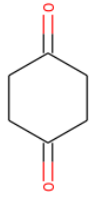
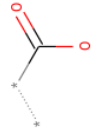
Natural product class	SMARTS (Additional requirements)	Visualization of SMARTS
peptides	<chem>CC(N([#1])C)=O</chem>	
	<chem>C[!r5;C](N([r5;C])C)=O</chem>	
phenylpropanoids	<chem>CC=Cc1ccccc1</chem>	
	<chem>C=CCc1ccccc1</chem>	
phenolics	<chem>[#1]O[!q3c]1[!q3c][!q3c][!q3c][!q3c]1</chem> (none of the carbons is part of another ringsystem)	
quinone	<chem>O=[#6]1C(=O)[#6]!-[#6][#6]=[#6]1</chem>	
	<chem>O=[#6]1[#6]!-[#6][#6](=O)[#6]!-[#6]1</chem>	
fatty acids ester	<chem>O=C(O)[!R;C][!R;C]</chem>	

Table A2: continued

Natural product class	SMARTS (Additional requirements)	Visualization of SMARTS
	<chem>O=C(O[!R;#6])[!R;C][!R;C]</chem>	
coumarine	<chem>O=C1C=Cc2ccccc2O1</chem>	
	<chem>O=C(OC=C1)c2c1ccccc2</chem>	
	<chem>OC1=CC(c2c(O1)ccccc2)=O</chem>	
depsidone	<chem>O=C1Oc2c(Oc3c1ccccc3)ccccc2</chem>	
	<chem>C1C2c(Oc3c1ccccc3)ccccc2</chem>	
sugar	<chem>OC1C(O)C(O)C(C(O)CO)O1</chem>	
	<chem>OC1C(O)C(O)C(O)C(CO)O1</chem>	
fluoren	<chem>C1c2c(c3c1ccccc3)ccccc2</chem>	

Table A2: continued

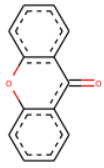
Natural product class	SMARTS (Additional requirements)	Visualization of SMARTS
xanthone	<chem>O=C1c2ccccc2Oc3c1ccccc3</chem>	
cyclic alkaloids	<chem>[n,N;R]</chem> (nitrogen is within a ringsystem)	<chem>N</chem>
acyclic alkaloids	<chem>[!R;N]</chem> (nitrogen is outside a ringsystem)	<chem>N</chem>

Table A3: Number of natural products belonging to a chemical class identified with SMARTS in MOE per Kingdom. One natural product can contribute to several chemical classes. ‘no_class’ represent natural products that were not matched with any defined SMART string.

	Plantae	Fungi	Animalia	Bacteria	Archaea	Protozoa	All
TotalNumber	18,025	2,486	5,295	3,546	92	426	25,698
NP_terpene	5,032	453	1,417	601	11	42	6,841
NP_phenylpropanoids	3,683	314	666	261	2	28	4,511
NP_alkaloid_cyclic	3,212	652	1,348	1,242	21	150	5,709
NP_fatty_acid_ester	2,544	395	768	682	26	88	3,932
NP_lacton	2,462	304	623	401	0	10	3,713
NP_sugar	2,422	106	324	127	8	25	2,880
NP_alkaloid_acyclic	2,167	584	1,276	1,508	32	187	4,629
NP_polyacetylene	1,236	232	485	345	2	31	2,115
NP_flavonoid	1,149	66	217	37	0	7	1,270
NP_steroid	912	99	183	42	0	2	1,170
NP_lactam	644	291	402	505	1	61	1,604
NP_benzofuran	631	26	124	26	0	2	746
no_class	585	203	285	297	24	57	943
NP_alkyl_chain	539	117	256	182	4	25	844
NP_macrocycle	431	154	381	428	2	20	1,298
NP_benzene	421	85	120	80	0	8	589
NP_peptoids	384	167	312	359	1	39	1,044
NP_quinone	381	44	68	80	0	1	544
NP_polyketide	309	79	106	92	0	14	510
NP_isoflavonoid	289	7	66	9	0	1	316
NP_coumarine	280	24	48	17	0	1	347
NP_xanthone	272	26	15	1	0	1	300
NP_phenanthrene	234	14	63	12	0	1	292
NP_anthracene	195	26	33	46	0	1	281
NP_peptides	160	103	157	194	0	17	503
NP_fluoren	6	1	2	1	0	0	15

Table A4: Taxonomic analysis of anti-infective effects. Plant families are given which were either in the analysis of all or of only native species above or below the 95 % interval, together with the total numbers of (anti-infective) species per family with the resulting ratio.

Family	all species				native species only			
	number of species	active species	ratio	95% CI	number of species	active species	ratio	95% CI
Acanthaceae	202	21	0.10	below	142	11	0.08	below
Acoraceae	2	2	1	above				
Actinidiaceae	18	0	0	below	18	0	0	below
Adoxaceae	11	0	0	below				
Amaryllidaceae	66	33	0.5	above				
Annonaceae					84	23	0.27	above
Apiaceae	22	18	0.82	above				
Apocynaceae	203	38	0.19	below	164	20	0.12	below
Aquifoliaceae	8	0	0	below				
Araceae	96	8	0.08	below	80	5	0.06	below
Araliaceae	50	5	0.1	below	40	2	0.05	below
Arecaceae	123	12	0.1	below	66	2	0.03	below
Asparagaceae	46	17	0.37	above				
Asphodelaceae	21	9	0.43	above				
Asteraceae	295	124	0.42	above				
Balsaminaceae	20	1	0.05	below	18	0	0	below
Begoniaceae	40	0	0	below	20	0	0	below
Berberidaceae	7	5	0.71	above				
Brassicaceae	31	15	0.48	above				
Bromeliaceae	36	2	0.06	below				
Burmanniaceae	9	0	0	below				
Calophyllaceae	11	6	0.55	above				
Caricaceae	2	2	1	above	1	1	1	above
Ceratophyllaceae	2	2	1	above	2	2	1	above
Chloranthaceae	5	3	0.6	above	3	2	0.67	above
Clusiaceae	16	9	0.56	above	15	8	0.53	above

Table A4: continued

Family	all species				native species only			
	number of species	active species	ratio	95% CI	number of species	active species	ratio	95% CI
Colchicaceae					3	2	0.67	above
Commelinaceae	39	1	0.03	below	28	0	0	below
Crypteroniaceae	1	1	1	above	1	1	1	above
Cupressaceae	12	11	0.92	above	1	1	1	above
Cyperaceae	218	8	0.04	below	188	5	0.03	below
Droseraceae	2	2	1	above	2	2	1	above
Elaeocarpaceae	22	2	0.09	below				
Ericaceae	47	4	0.09	below	41	1	0.02	below
Fabaceae	553	201	0.36	above	341	90	0.26	above
Fagaceae	26	2	0.08	below	24	0	0	below
Garryaceae	1	1	1	above				
Gesneriaceae	71	3	0.04	below	59	2	0.03	below
Hernandiaceae	5	3	0.6	above	5	3	0.6	above
Hydrocharitaceae	20	0	0	below	17	0	0	below
Hypericaceae	9	5	0.56	above	8	5	0.63	above
Hypoxidaceae	4	3	0.75	above	4	3	0.75	above
Lamiaceae	157	63	0.4	above	99	32	0.32	above
Lentibulariaceae	9	0	0	below				
Linderniaceae	21	1	0.05	below	18	0	0	below
Loranthaceae	32	4	0.13	below				
Lythraceae					20	7	0.35	above
Magnoliaceae	17	8	0.47	above				
Marantaceae	12	0	0	below				
Melastomataceae	99	3	0.03	below	91	1	0.01	below
Meliaceae	67	29	0.43	above	61	24	0.4	above
Menispermaceae	42	16	0.38	above	40	15	0.38	above
Molluginaceae	3	3	1	above	3	3	1	above
Moringaceae	2	2	1	above				
Muntingiaceae	1	1	1	above				

Table A4: continued

Family	all species				native species only			
	number of species	active species	ratio	95% CI	number of species	active species	ratio	95% CI
Nelumbonaceae	1	1	1	above				
Onagraceae	24	2	0.08	below				
Orchidaceae	1016	39	0.04	below	745	20	0.03	below
Orobanchaceae	17	1	0.06	below				
Papaveraceae	11	7	0.64	above				
Philydraceae	1	1	1	above				
Phyllanthaceae	94	17	0.18	below				
Phytolaccaceae	6	4	0.67	above	1	1	1	above
Poaceae	373	34	0.09	below	276	9	0.03	below
Podocarpaceae	6	5	0.83	above	5	4	0.8	above
Polygalaceae	28	3	0.11	below				
Polygonaceae	44	18	0.41	above	28	10	0.36	above
Primulaceae	85	11	0.13	below	77	8	0.1	below
Rhizophoraceae	11	8	0.73	above	9	6	0.67	above
Rubiaceae	344	61	0.18	below	282	38	0.14	below
Ruppiaceae	1	1	1	above	1	1	1	above
Rutaceae	82	45	0.55	above	69	33	0.48	above
Saururaceae	1	1	1	above				
Saxifragaceae	2	2	1	above	1	1	1	above
Simaroubaceae	7	5	0.71	above	6	4	0.67	above
Solanaceae	83	33	0.4	above				
Tamaricaceae	1	1	1	above				
Typhaceae	2	2	1	above	2	2	1	above
Urticaceae	97	5	0.05	below	92	5	0.05	below
Violaceae	20	1	0.05	below	18	0	0	below
Vitaceae	64	4	0.06	below	60	2	0.03	below

Table A5: Environmental variables used for the species distribution modelling together with the mean relative contribution to the models and its standard deviation.

Source	Variable	Mean	SD
WorldClim	Isothermality (Mean diurnal range/Temperature annual range)	8.27	4.77
WorldClim	Precipitation of Warmest Quarter	7.40	3.98
WorldClim	Precipitation of Coldest Quarter	7.10	3.65
WorldClim	Precipitation of Wettest Quarter	7.06	3.65
SoilGrids	Derived available soil water capacity (volumetric fraction) with FC = pF 2.0	7.00	3.86
SoilGrids	Depth to bedrock (R horizon) up to 200 cm	6.91	3.98
SoilGrids	Soil pH x 10 in KCl	6.90	3.80
WorldClim	Precipitation Seasonality (Coefficient of Variation)	6.68	3.26
SoilGrids	Cummulative probability of organic soil based on the TAXOUSDA and TAXNWRB	6.62	2.98
SoilGrids	Absolute depth to bedrock (in cm)	6.59	3.16
SoilGrids	Cation exchange capacity of soil in cmolc/kg	6.41	3.14
SoilGrids	Coarse fragments volumetric in %	5.90	2.92
SoilGrids	Soil organic carbon density in kg per cubic-m	5.70	3.33

Table A6: Trigger (“ISO”) and negation (“NEG”) words used in the text mining process to identify natural product - natural sources relations and to prevent false postive extraction, respectively.

ISO	NEG
abundant	absent
common	accumulate

ISO	NEG
component	accumulated
compose	accumulates
composed	accumulating
composes	admit
composing	admits
comprise	admitted
comprised	admitting
comprises	adsorbed
comprising	adsorbed
consist	adsorbes
consisted	adsorbing
consisting	although
consists	analog
constituent	antibodies
contain	antibody
contained	applied
containing	applies
contains	apply
content	applying
cover	area
covered	artificial
covering	assay
covers	assimilate
demonstrate	assimilated
demonstrated	assimilates
demonstrates	assimilating
demonstrating	association
derive	biotransformation
derived	cell culture
derives	cell cultures
deriving	cell line
detached	cell lines
detected	cell suspension culture

ISO	NEG
determine	cell suspension cultures
determined	clone
determines	commercial
determining	commercially
discovered	compare
distributed	control
dominate	control group
dominated	control groups
dominates	controls
dominating	could
embodied	culture
embodies	cyclase
embody	degrade
embodying	degraded
embrace	degrading
embraced	dehydrogenase
embraces	district
embracing	effect
encompass	effect
encompassed	effects
encompasses	engineer
encompassing	engineered
enhance	engineering
enhanced	engineers
enhances	enzyme
enhancing	equivalence
enriched	equivalent
explain	equivalents
explained	ethnic group
explaining	ethnic groups
explains	except
extracted	expression
found	extraction

ISO	NEG
gained	fed
identification	feed
identified	feeding
include	feeds
included	fermentation
includes	fermented
including	free
incorporate	fumigant
incorporated	gene
incorporates	genes
incorporating	growth
involve	herbicide
involved	heterologus
involves	incorporate
involving	incorporated
isolated	incorporates
isolation	incorporating
known	infect
major active compound	infected
major component	infecting
major components	infection
major compound	inhibition
major quantities	ligand
minor compound	medium
minor compounds	metabolize
minor quantities	metabolized
observed	metabolizing
obtained	mitogen
originate	mixture
originated	mutagen
originates	mutagenic
originating	mutant
possessing	n.a.

ISO	NEG
possesse	n.d.
possessed	na
possesses	nd
predominate	neither
predominated	never
predominates	not
predominating	not detected
presence	phase
present	poison
produce	poisoned
produced	poisoning
produces	potential
producing	protein
production	quantifications
proved	region
release	regulated
released	resistance
releases	scavenging
releasing	standard
reported	stimulation
reveal	synthase
revealed	synthesized
revealing	synthetic
reveals	synthetized
segregated	titration
separated	trait
show	transcript
showed	transcription
showing	transformation
shows	transformed
split	transgenic
subsume	treated
subsumed	treating

ISO	NEG
subsumes	treatments
subsuming	would
uncovered	
yield	
yielded	
yielding	
yields	

Table A7: ChEMBL Bioassays with the top ten of most matched natural products.

Assay-Id	Number of NPs	Description
1265865	767	Inhibition sodium fluorescein uptake in OATP1B3 cells
1265862	759	Inhibition sodium fluorescein uptake in OATP1B1 cells
954272	726	Inhibitors of human tyrosyl-DNA phosphodiesterase 1
954305	725	Inhibitors of human tyrosyl-DNA phosphodiesterase 1
1301778	681	Identifying genotoxic compounds (cytotoxicity against isogenic chicken DT40 cell lines) – Rev3 mutant cell line
736784	673	multi-species drug-induced liver injury (DILI) dataset, effect in humans
736785	673	multi-species drug-induced liver injury (DILI) dataset, effect in rodents
736786	673	multi-species drug-induced liver injury (DILI) dataset, effect in non-rodents

Assay-Id	Number of NPs	Description
1301858	652	Identifying genotoxic compounds (cytotoxicity against isogenic chicken DT40 cell lines) – wild type cell line
1301573	617	Disruption of mitochondrial membrane potential

Table A8: ChEMBL targets with the top ten and number 15 and 22 (Position in Ranking) of most matched natural products targets.

Target-Id	Number of NPs	Preferred Name	Number of assays for target	Position in Ranking
22226	3,943	Unchecked	35,320	1
22224	1,643	ADMET	6,088	2
50425	1,161	Plasmodium falciparum	2,745	3
22229	1,158	No relevant target	1,261	4
50594	1,134	Mus musculus	15,052	5
50597	1,103	Rattus norvegicus	193,381	6
80682	1,093	A549	2,185	7
80224	1,088	MCF7	2,685	8
17045	1,068	Cytochrome P450 3A4	369	9
11365	903	Cytochrome P450 2D6	144	10
50185	874	Staphylococcus aureus	12,478	15
50212	749	Escherichia coli	6,909	22

Table A9: ChemOnt classes which matched at least 100 natural products together with their respective ratio of matched natural products.

	Ratio [%]
Lipids and lipid-like molecules	33.4
Prenol lipids	23

	Ratio [%]
Phenylpropanoids and polyketides	18.4
Organoheterocyclic compounds	16
Benzenoids	10
Organic oxygen compounds	7.5
Organooxygen compounds	7.4
Flavonoids	6.9
Steroids and steroid derivatives	5.7
Diterpenoids	5
Alkaloids and derivatives	4.9
Triterpenoids	4.9
Benzene and substituted derivatives	4.8
Organic acids and derivatives	4.8
Terpene glycosides	4.4
Fatty Acyls	4.2
Carbohydrates and carbohydrate conjugates	3.9
Carboxylic acids and derivatives	3.8
Benzopyrans	3.4
Sesquiterpenoids	3.3
1-benzopyrans	3.2
Terpene lactones	3
Triterpene glycosides	2.5
Triterpene saponins	2.5
Glycosyl compounds	2.4
Dibenzopyrans	2.2
Xanthenes	2.2
Amino acids, peptides, and analogues	2.2
Lignans, neolignans and related compounds	2.1
Carbonyl compounds	2.1

	Ratio [%]
Flavonoid glycosides	2
Monoterpenoids	1.9
Xanthones	1.9
Isoflavonoids	1.9
Cinnamic acids and derivatives	1.8
Flavonoid O-glycosides	1.8
Indoles and derivatives	1.7
O-methylated flavonoids	1.6
Ketones	1.6
Steroidal glycosides	1.6
Phenols	1.5
Amino acids and derivatives	1.5
Hydroxycinnamic acids and derivatives	1.4
Flavans	1.3
2-arylbenzofuran flavonoids	1.3
Benzoic acids and derivatives	1.3
Alpha amino acids and derivatives	1.3
Coumarins and derivatives	1.3
Fatty alcohols	1.3
Steroid lactones	1.3
Phenolic glycosides	1.2
Kaurane diterpenoids	1.1
Flavones	1.1
Tannins	1.1
Quinolines and derivatives	1.1
Anthracenes	1.1
Steroidal saponins	1.1
Hydrolyzable tannins	1
Diarylheptanoids	1
Limonoids	1
Linear 1,3-diarylpropanoids	1

	Ratio [%]
O-glycosyl compounds	1
Linear diarylheptanoids	0.9
Sesquiterpene lactones	0.9
Coumaric acids and derivatives	0.9
Alcohols and polyols	0.9
Phenanthrenes and derivatives	0.8
Naphthalenes	0.8
Anthraquinones	0.8
Chalcones and dihydrochalcones	0.8
Diterpene lactones	0.8
Stilbenes	0.8
Aryl ketones	0.8
Lactones	0.8
Oligosaccharides	0.8
Organonitrogen compounds	0.7
Organic nitrogen compounds	0.7
Fatty acids and conjugates	0.7
Furanoid lignans	0.7
7-O-methylated flavonoids	0.7
Iridoid O-glycosides	0.7
Peptides	0.7
Bicyclic monoterpenoids	0.7
Phenylketones	0.7
Methoxyphenols	0.7
Aporphines	0.7
Alkyl-phenylketones	0.7
Colensane and clerodane diterpenoids	0.6
Fatty acyl glycosides	0.6
Fatty acyl glycosides of mono- and disaccharides	0.6
Carbazoles	0.6

	Ratio [%]
Cyclic ketones	0.6

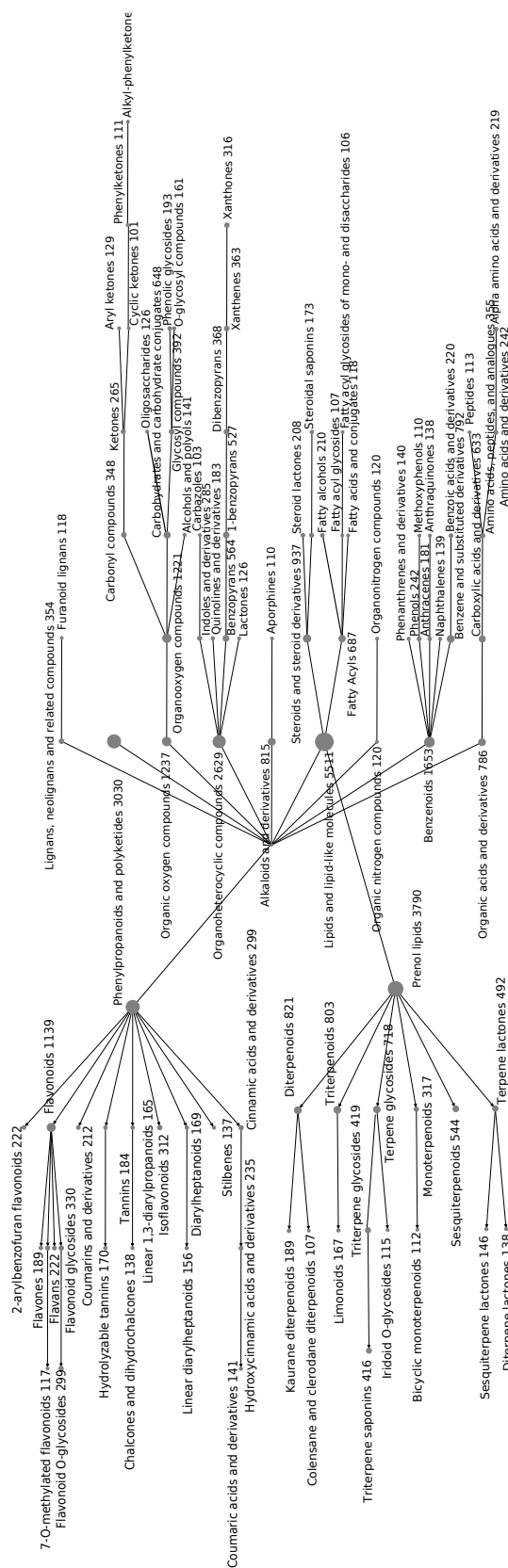


Figure A1: Tree representation of ChemOnt ontology¹⁰⁰ with all classes which at least matched 100 natural products from our data set. The bigger the node, the more natural products matched the respective class.

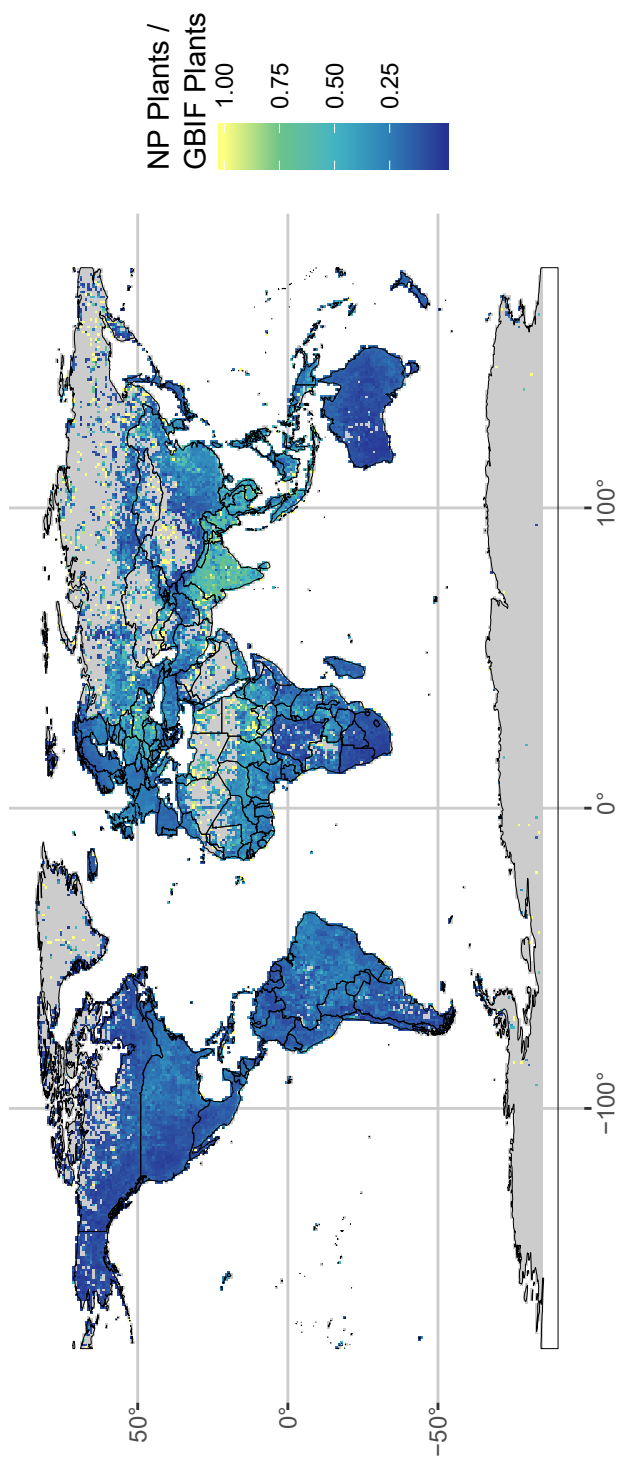


Figure A2: Ratio of plants with natural product information extracted from literature (NP Plants) to GBIF plants. Grid cells are of approximately 100km x 100km and each plant counted uniquely if an occurrence of it can be found in the grid cell.

Curriculum vitae

Name	Anne-Kahtrin Hartig, geb. Blume
seit 09/2019	Wissenschaftliche Mitarbeiterin im Datenintegrationszentrum des Universitätsklinikum Halle (Saale)
01/2016-07/2019	Promotion am Leibniz-Institut für Pflanzenbiochemie in Halle (Saale), Abteilung Natur- und Wirkstoffchemie von Prof. L. A. Wessjohann, AG Computerchemie von PD Dr. Brandt Thema: "Chemoinformatical Analysis of Natural Products"
10/2013-12/2015	Master Biochemie an der Martin-Luther-Universität Halle-Wittenberg Abschluss: Master of Science Masterarbeit: "Characterization of adenylate isopentenyl transerases from <i>Arabidopsis thaliana</i> based on protein homology models, virtual screening and <i>in vitro</i> studies"
10/2010 - 09/2013	Bachelor Biochemie an der Martin-Luther-Universität Halle-Wittenberg Abschluss Bachelor of Science Bachelorarbeit: "Strukturelle und funktionelle Charakterisierung der Variante A287G des Enzyms Pyruvatdecarboxylase aus <i>Saccharomyces cerevisiae</i> "
08/2008 - 06/2010	Abitur Graf Münster Gymnasium Bayreuth

Erklärungen

1. Ich erkläre, dass ich mich an keiner anderen Hochschule einem Promotionsverfahren unterzogen bzw. eine Promotion begonnen habe.
2. Ich erkläre, die Angaben wahrheitsgemäß gemacht und die wissenschaftliche Arbeit an keiner anderen wissenschaftlichen Einrichtung zur Erlangung eines akademischen Grades eingereicht zu haben.
3. Ich erkläre an Eides statt, dass ich die Arbeit selbstständig und ohne fremde Hilfe verfasst habe. Alle Regeln der guten wissenschaftlichen Praxis wurden eingehalten; es wurden keine anderen als die von mir angegebenen Quellen und Hilfsmittel benutzt und die den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht.

Datum und Unterschrift