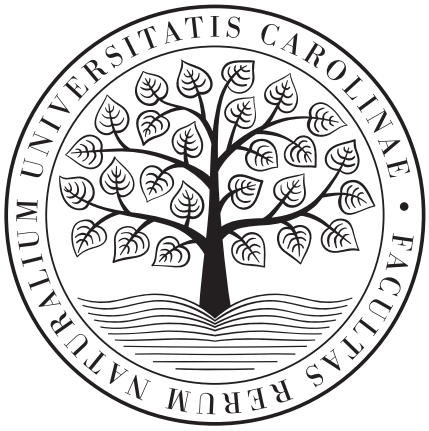
Charles University in Prague
Faculty of Science

BACHELOR THESIS



Natália Komorníková

Development and analysis of a database of reactions catalyzed by cytochrome P450 enzymes for machine learning applications

Vytvoření a analýza databáze reakcí katalyzovaných cytochromy P450 pro strojové učení

Department of Cell Biology

Supervisor of the bachelor thesis: Mgr. Tomáš Pluskal, Ph.D.

Study programme: Bioinformatics

Study branch: Bioinformatics

Prague 2023

I declare that I carried out this bachelor thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In ..... date ....

Author's signature

I want to express my gratitude to my supervisor Mgr. Tomáš Pluskal, Ph.D., for his guidance, valuable advice, expertise, immediate feedback and the time he devoted to me. My colleagues' comments and questions, especially to Michal, whose scripts immensely helped me in my work.

A big thank you belongs to my family for their unconditional love and support throughout my studies.

A rightfully exhausting thank you belongs to my friends, especially Paty and Baška, for always cheering me up and being supportive.

Last but not least, I would like to thank Aďko for his constant support and his stress relief capabilities, which kept me in my healthy mind during the last months.

Title: Development and analysis of a database of reactions catalyzed by cytochrome P450 enzymes for machine learning applications

Author: Natália Komorníková

Department: Department of Cell Biology

Supervisor: Mgr. Tomáš Pluskal, Ph.D.,

Abstract: Cytochrome P450 enzymes are hemoproteins showing extraordinary diversity in the reactions they catalyze. We developed a database containing all the needed data to provide a comprehensive data source on reactions catalyzed by cytochrome P450 enzymes. This data mainly includes information about the substrates, products of characterized reactions, and the sequence of these enzymes. The database was developed by collecting data from reliable protein and reaction databases like UniProt and RHEA. The work presents an in-depth analysis of the created database of reactions catalyzed by cytochrome P450 enzymes. This database can be utilized for future machine learning approaches to predict the function of uncharacterized cytochrome P450s.

Keywords: cytochrome P450; database; substrate; product; catalyst; data mining; distribution of data; UniProt; RHEA

Abstrakt: Cytochromy P450 jsou hemoproteiny vykazující mimořádnou rozmanitost reakcí, které katalyzují. Vyvinuli jsme databázi obsahující všechny potřebné údaje, abychom poskytli komplexní zdroj dat o reakcích katalyzovaných enzymy cytochromu P450. Tato data zahrnují především informace o substrátech, produktech charakterizovaných reakcí a sekvenci těchto enzymů. Databáze byla vytvořena shromážděním údajů ze spolehlivých databází proteinů a reakcí, jako jsou UniProt a RHEA. Práce představuje hloubkovou analýzu vytvořené databáze reakcí katalyzovaných enzymy cytochromu P450. Tato databáze může být v budoucnu využita pro přístupy strojového učení k předpovídání funkce necharakterizovaných cytochromů P450.

Klíčova slova: cytochrom P450; databáze; substrát; produkt; katalyzátor; data mining; distribuce dat; UniProt; RHEA

Contents

Introduction

This thesis aims to contribute to research on cytochrome P450 enzymes initiated by my supervisor Mgr. Tomáš Pluskal, Ph.D. Cytochrome P450 enzymes are a significant class of enzymes metabolizing drugs, toxins, and endogenous compounds in all living organisms. They are also involved in the biosynthesis of various compounds, such as hormones, fatty acids, and secondary metabolites. However, the diverse substrates they can bind to and the complex nature of their catalytic activity make it difficult to predict their substrate specificity and the reaction outcomes. Understanding the function and regulation of cytochrome P450s is crucial in various fields, including drug discovery and determining a potential risk associated with exposure to toxic substances.

This thesis has three main objectives. Firstly, we must address the limited availability of high-quality data on reactions catalyzed by cytochrome P450 enzymes. The existing databases of cytochrome P450 enzymes include: The Cytochrome P450 Homepage [1], which is a precious source for nomenclature and sequence information, and the Plant CYPs database[1], which contains only plant P450s with information like clan, family, species and the function. These databases do not contain any information on reactions these enzymes catalyze, nor the information needed for machine learning applications. Therefore, the main aim of this thesis is to collect data and curate it to create a comprehensive database of all known annotated reactions catalyzed by cytochromes P450. The database we aim to develop can be used as a training and validation set for machine learning models in the future. By curating and standardizing data from various sources, the database will help to advance the development of a machine learning-based approach for predicting the activity of uncharacterized cytochrome P450s.

We collected the data for our database from various sources such as UniProt, RHEA, and ChEBI. After collecting the data, we analyzed the database to see how the characterized reactions are distributed in the kingdoms of organisms and how long these enzymes typically are. We also wanted to see the redundancy of the substrates in the different kingdoms by visualizing the chemical space and the most common classes of substrates of these reactions.

[1]The Plant Cytochrome P450 Database

In the first chapter, we will look at the cytochrome P450 enzymes, their catalytic cycle, the typical reactions they catalyze and their applications in biotechnology and industry. Next, we will show how we collected the data, how we organized them, the criteria we used for including them in the database and the overall structure of the database. Finally, we will broadly analyze the created database.

Chapter 1

Cytochrome P450 enzymes

Cytochrome P450 monooxygenases, also known as CYPs or P450s, are hemoproteins widely distributed in all types of organisms and play essential roles in natural product biosynthesis, xenobiotic degradation, steroid biosynthesis, and drug metabolism [2, 3].

Cytochromes P450 can perform oxidative reactions, with a qualitative distinction between prokaryotic and eukaryotic CYPs. Eukaryotic CYPs are generally membrane-bound proteins, unlike prokaryotic CYPs [4, 5]. In eukaryotic cells, the reactions CYPs catalyze occur in the endoplasmic reticulum. To function properly, cytochrome P450s require reducing cofactors like NAD(P)H and redox partner proteins [4].

The name "cytochrome P450" comes from these enzymes' peak absorbance at 450 nm when bound to carbon monoxide [6]. This property gives the designation "P450" or "pigment-450 nm."

They are known for their exceptional versatility in catalyzing a wide range of reactions on various substrates, making them one of the most versatile biocatalysts in nature. Cytochrome P450 enzymes facilitate the transfer of an oxygen atom from molecular oxygen to different biological substrates. The second oxygen atom is converted into a water molecule via reduction by two electrons [7]. We can describe the oxidation reaction catalyzed by a CYP by the equation:

$$\mathbf{RH + O_2 + NAD(P)H + H^+ \xrightarrow{P450} R{-}OH + NAD(P)^+ + H_2O}.$$

The substrates are on the left-hand side, and the products are on the right-hand side of the reaction equation.

The cellular location and domain architecture of CYPs categorize them into ten different classes. Researchers further divide these classes into families based solely on the homology of their amino acid sequence [6].

1.1 The nomenclature of CYPs

CYPs are named based on a standardized nomenclature system that assigns names to the family of genes. The nomenclature system is based on the amino acid sequence identity [8]. A minimum of 40% amino acid sequence identity is required for assignment to the same family cluster. Additionally, more than 55% sequence identity is required for subfamily members. CYPs follow the nomenclature: CYP, followed by an Arabic numeral (family), a letter (subfamily), and another Arabic numeral (individual gene) [8, 9]. For example, CYP71A1 is a member of the CYP71 family, subfamily A, and is the first gene in that family to be identified.

While CYPs are named based on a standardized nomenclature system that assigns names to the family of genes, our work does not use this organization of enzymes. Instead, we focus on the function of these enzymes, as our research is centered around developing machine-learning applications for predicting the functions of uncharacterized CYPs. Therefore, our approach emphasizes the functional characterization of CYPs rather than their classification based on amino acid sequence identity.

1.2 Architecture and the Functional Domains of Cytochrome P450 Enzymes: IPR036396 and PF00067

There is a sufficient number of different CYPs tertiary structures to state that the P450-fold is conservative [4], meaning the fold of these enzymes remains by large the same, while the precise positioning of different structural elements may vary significantly. The typical cytochrome P450 structure is shown in Figure 1.1. Other enzymes exhibit the P450-fold but do not catalyze traditional P450 chemistry. For example, the NO reductase, P450nor, prostacyclin synthase and others [4].

1.2.1 Functional Domains

The cytochrome P450 superfamily (**IPR036396**[1]) is a large and diverse group of enzymes metabolizing various endogenous and exogenous compounds.

A commonly recognized protein family associated with this superfamily is the **PF00067** or **IPR001128**[2], essential for binding the heme cofactor required for

[1]The InterPro entry for cytochrome P450 superfamily
[2]The InterPro entry for the Pfam domain

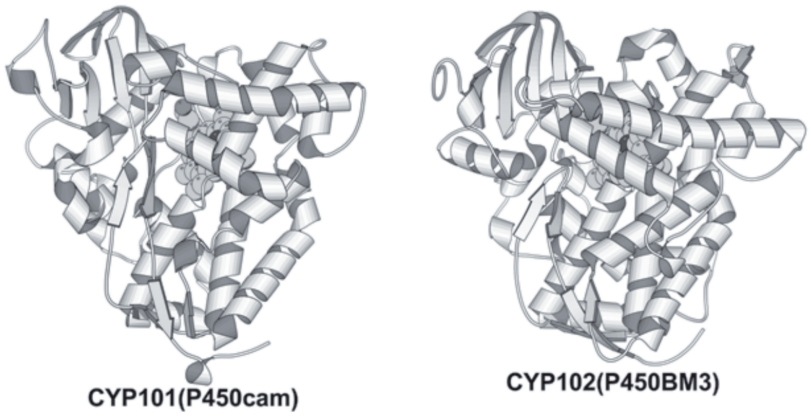


Figure 1.1 Structures of two CYPs illustrating the common P450-fold. Figure adopted from [4].

catalytic activity [9]. In the InterPro database[3] we can also find other overlapping entries for the cytochrome P450 superfamily, like **IPR002397** (the Cytochrome P450, B-class), **IPR002399** (the Cytochrome P450, mitochondrial) and others. However, if we look at the domain architectures[4] for the cytochrome P450s, there are overall 3603 architectures found, where each of these architectures contains the PF00067 domain. This is important for the data collection process used in 2.1.2.

1.3 Promiscuity of CYPs

There are two types of cytochrome P450 enzymes: substrate-specific and highly promiscuous [10].

Some CYPs are highly specific for particular substrates and only metabolize a narrow range of compounds, while the promiscuous ones can metabolize a wide range of diverse substrates. These substrates include drugs, toxins, and endogenous compounds. Promiscuity arises from the active site's architecture and the heme iron center's flexibility, which can accommodate various chemical structures.

Plant CYPs were initially thought to be highly substrate-specific, regioselective[5] and stereoselective[6] enzymes [10]. However, recent functional data counter

[3]A bioinformatics database that provides functional analysis of protein sequences by classifying them into families, domains, and functional sites.

[4]Domain architectures found in the InterPro database. We are talking about PFAM domains (the InterPro database shows both the PFAM and the InterPro entry for each domain)

[5]Ability to catalyze reactions at specific positions within a substrate mol. selectively

[6]Ability to selectively catalyze reactions producing specific stereoisomers of a substrate mol.

this belief. Werck-Reichhart suggest that the ability to catalyze multiple reactions with different substrates may be a driving force behind the evolution of CYPs in plants [10].

In plants, highly promiscuous CYPs have been reported mainly in terpenoid pathways [10]. Different plant P450 families have varying degrees of substrate specificity and permissiveness. Highly promiscuous plant CYPs tend to retain gene duplicates and undergo evolutionary expansion.

1.4 The cytochrome P450 catalytic cycle

The catalytic cycle of CYPs involves multiple complex steps with the formation of many different transient intermediates.

The P450 catalytic system includes the substrate, a P450 enzyme (substrate binding and oxidative catalysis), the redox partner (electron transfer shuttle), and the cofactor (NAD(P)H) [2]. The enzymes mediate the catalytic oxidation of target substances, whereby the substrate initially binds to the active site within the heme domain [6].

There are five stages in the catalytic cycle of CYPs: substrate binding, first reduction, oxygen binding, second reduction and product release [11]. Generally, the ferric resting state of the enzyme accepts an RH substrate and forms a hydroxylated product R-OH, which is then released from the active site. After forming the product, a water molecule binds and restores the resting state [2]. Figure 1.2 represents the P450 catalytic cycle with its five stages.

To activate oxygen in P450 enzymes, two protons and two electrons must be delivered. The electrons typically come from NAD(P)H and are supplied separately [12]. Electrons need to be introduced to the heme-iron continuously to preserve the P450 catalytic cycle.

1.5 Basic reactions catalyzed by cytochrome P450 enzymes

While all cytochrome P450 enzymes share a common mechanism of catalysis, the substrate specificity and catalytic activity can vary significantly between different organisms. Notably, plant cytochromes have specific reaction attributes compared to other organisms [13].

Plant CYPs participate in various biochemical pathways to produce a vast diversity of natural plant products, including secondary metabolites such as phenylpropanoids, alkaloids, terpenoids, lipids, cyanogenic glycosides, glucosinolates, and plant hormones [13]. These secondary metabolites are not produced

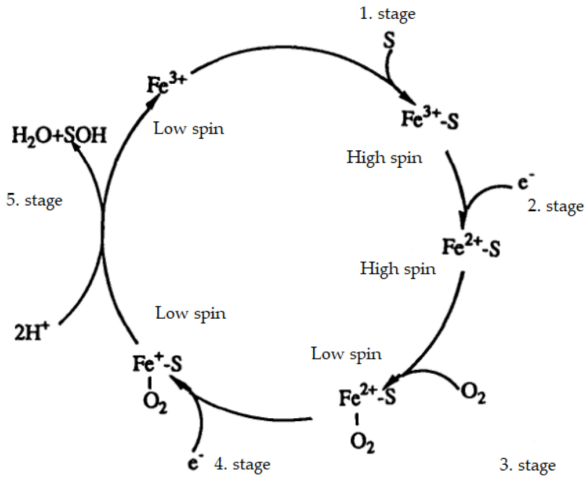


Figure 1.2 Schematic representation of the catalytic cycle of a P450 enzyme and its five stages. Here the substrate is represented by **S**. Adapted figure from [11].

in animals [13] and contribute significantly to the phytochemical diversity via various oxidative modifications of the carbon skeleton.

Additionally, some plant cytochromes may have specific roles in detoxifying herbicides [14] and other environmental toxins which are not present in animal systems. On the other hand, animal cytochromes may play a more prominent role in the metabolism of endogenous hormones and the detoxification of xenobiotics, such as drugs and environmental pollutants.

The most common reaction catalyzed by CYPs is hydroxylation, where an oxygen atom is added to a carbon of the substrate (an organic molecule), producing an alcohol. Additionally, CYPs can catalyze reactions such as dealkylation, dehalogenation, and epoxidation [15]. These reactions involve introducing oxygen or removing other atoms from the substrate, such as nitrogen, sulfur, or halogens. CYPs can also perform several reactions on heterocyclic compounds containing a ring structure with at least one non-carbon atom [4]. We will introduce the plant-specific reactions CYPs catalyze in section 1.6.

The most fundamental reactions we will introduce here are carbon hydroxylation, heteroatom oxygenation, heteroatom release, and epoxidation.

1.5.1 Carbon hydroxylation

Carbon hydroxylation, or C-oxidation, is a widely occurring chemical reaction. It results in the formation of alcohols from various substrates, such as steroids and alkanes [16]. The way cytochrome P450 catalyzes the hydroxylation of carbons

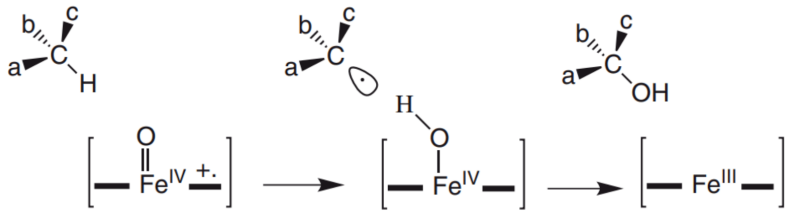


Figure 1.3 A figure showing the carbon hydroxylation. Figure adopted from [4].

involves the removal of a hydrogen atom, along with its electron, from a C-H bond by the compound I ferryl species. This process produces two entities: a substrate carbon radical and a complex of ferric iron with a hydroxyl radical. In the next step, the hydroxyl radical reacts with the substrate carbon radical to form a hydroxylated product while also regenerating the ferric enzyme [4]. The reaction is shown in Figure 1.3.

1.5.2 Heteroatom oxygenation

Oxygenation is the process of adding an oxygen atom to a molecule. CYPs are versatile enzymes that can add oxygen to heteroatoms such as nitrogen, sulfur, phosphorus, and iodine atoms. Many N and S oxygenations were previously attributed to flavin-containing monooxygenase, but they are P450 reactions [16]. Cytochrome P450 enzymes catalyze oxidations in which the ferryl oxygen is added to the nitrogen instead of an adjacent carbon in the metabolic product [4]. The reaction is shown in Figure 1.4.

1.5.3 Heteroatom release (dealkylation)

These transformations involve adding a hydroxyl group on a carbon next to the heteroatom and eliminating the heteroatom by creating a carbonyl group [4]. It results in an unstable intermediate that subsequently undergoes cleavage between the carbon and the heteroatom, leading to compounds such as carbinolamines, acetals, thioacetals, and gemhalohydrins [16]. The mechanism of this reaction can be described as a typical carbon hydroxylation, which may occur similarly to halogens like fluorides, chlorides, and possibly bromides, as evidenced by high kinetic hydrogen isotope effects [16]. The reaction is shown in Figure 1.4.

1.5.4 Epoxidation

Epoxidation reactions involve the formation of an epoxide functional group, which consists of a three-membered ring containing an oxygen atom and two

adjacent carbon atoms [12]. CYPs commonly catalyze epoxidations. Cytochrome P450 enzymes efficiently epoxidize olefins with retention of stereochemistry in the cis-olefin. However, the P450-catalyzed epoxidation of some olefins produces side products besides the corresponding epoxides. Terminal olefins can cause the inhibition of cytochrome P450 by forming an irreversible porphyrin N-alkylation product. The new carbon-nitrogen bond is formed with the terminal carbon atom C1 of the olefin [17, 16]. Epoxidation is of interest in toxicology, organic synthesis, and chemists studying the mechanism of chemical epoxidation. The reaction is shown in Figure 1.4.

1.6 Plant-specialized Cytochrome P450 reactions

One of the most extensive gene superfamilies in plant genomes is the plant CYPs [13]. The plant genome contains as much as 1% of CYPs in their complete gene annotations, suggesting that plants are enormous sources for P450-dependent reactions [13]. We will show unusual P450-dependent reactions in plant secondary metabolisms because they differ from the reactions CYPs typically catalyze in other organisms. These reactions include a three-step oxidation, Baeyer-Villiger oxidation, methylenedioxy-bridge formation and sterol desaturation. The reactions are shown in Figure 1.5.

1.6.1 Three-step oxidation

The oxidative conversion of *ent*-kaurene to gibberellin A_{12} accompanies two CYPs. Firstly, CYP701A [13] catalyzes the three-step oxidation of *ent*-kaurene to *ent*-kaurenoic acid. The *ent*-kaurenoic acid is then oxidized to GA_{12} by CYP88A [13]. Both oxidations are shown in 1.5.

Gibberellins are plant hormones derived from diterpenoids [13]. They play a crucial role in regulating seed germination and the growth of plants.

1.6.2 Methylenedioxy-bridge formation

One of the main furofuran lignans in sesame seeds is (+)-sesamin, which contains two unique methylenedioxy bridges [13]. Research has demonstrated that the sesamum CYP81Q1 enzyme [13, 19] is involved in the biosynthesis of sesamin and that it catalyzes the creation of the two methylenedioxy bridges on (+)-pinoresinol, resulting in the production of (+)-sesamin via (+)-piperitol. The reaction is shown in Figure 1.5.

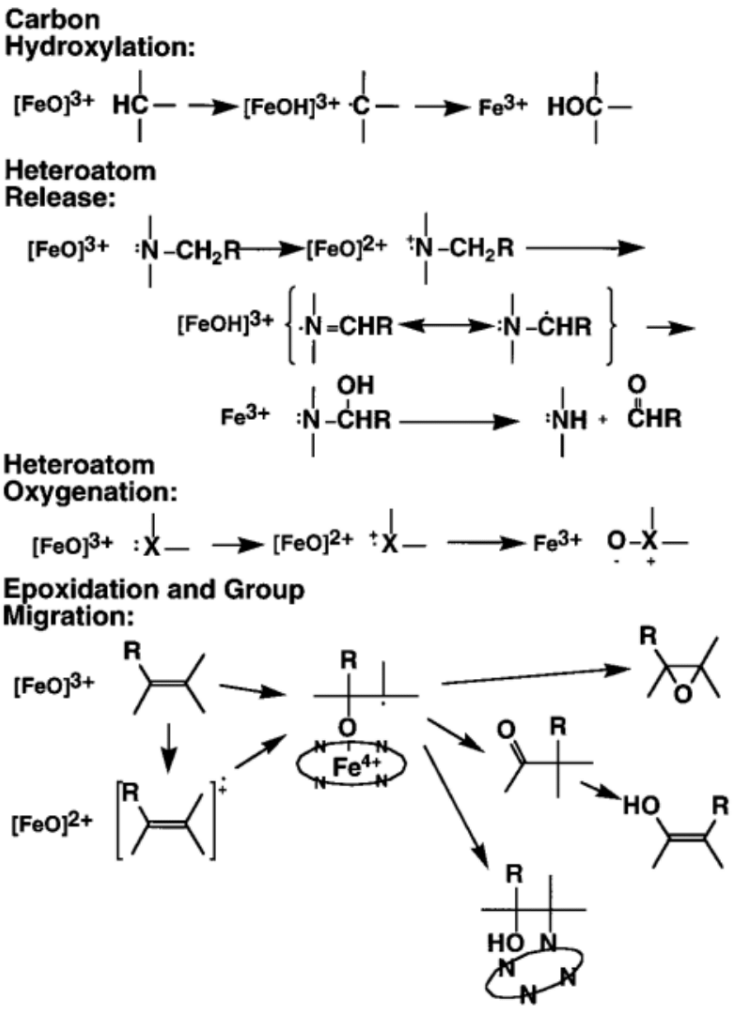


Figure 1.4 A figure showing schematically the basic reactions catalyzed by Cytochrome P450 enzymes. Figure adopted from [16].

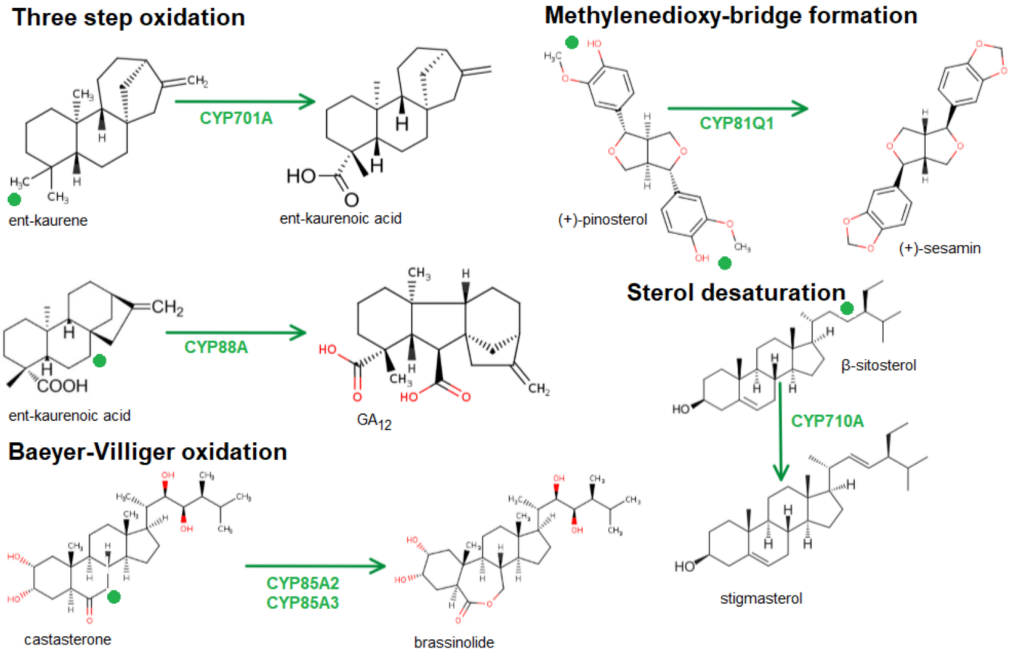


Figure 1.5 P450-dependent reactions in plant secondary metabolisms. Figure adapted from [18].

1.6.3 Baeyer-Villiger oxidation

The product of Baeyer-Villiger oxidation is brassinolide. The biosynthesis of brassinosteroids involves several P450-dependent oxygenations. Brassinolide is biologically the most active compound of brassinosteroids that functions in plant growth regulation and development [13]. The shown reaction in Figure 1.5 is catalyzed by brassinolide synthase: CYP85A2.

1.6.4 Sterol desaturation

Stigmasterol and brassicasterol are $\Delta 22$-unsaturated sterols found in plants and fungi. They contain a double bond at the C-22 position in the side chain. CYP710A, a sterol C-22 desaturase, has been identified to catalyze the desaturation of β-sitosterol to stigmasterol in plants [13]. This C-22 desaturation is shown in Figure 1.5.

1.7 Applications of CYPs in biotechnology and industry

CYPs have a unique advantage over other enzymes because no other enzyme group has a similar space of accepted substrates and catalyzed reaction types [20]. Cytochrome P450s offer a wide range of applications in biotechnology. They include the synthesis of pharmaceuticals, agrochemicals, and high-value chemicals. CYPs can also be used for bioremediation by helping the degradation of environmental pollutants. Additionally, they can potentially be used in biosensors and biofuel cells. These benefits are limited by their low stability, reduced activity in certain conditions, narrow substrate scope, and dependence on cofactor and redox partner [2].

Chapter 2

The database of reactions catalyzed by cytochrome P450 enzymes

The practical part of this thesis consists of developing a database of reactions catalyzed by cytochrome P450 enzymes for machine-learning applications. To carry out meaningful analyses and provide accurate conclusions, it is essential to ensure that the data is of high quality, comprehensive and relevant. This chapter provides an overview of the data collection, the sources of data and the curation processes used.

2.1 Collection of data from other databases

We worked with several sources of data, mainly with the UniProt database.

UniProt The Universal Protein Resource (UniProt) [21] is a high-quality, comprehensive and freely accessible database of protein sequence and functional information. It provides an up-to-date, complete body of protein information. The UniProt databases are UniProtKB (UniProt Knowledgebase), UniRef (UniProt Reference Clusters), and UniParc(the UniProt Archive). UniProt also provides cross-references to other relevant databases we used in the developing process.

RHEA Reactome Homology Enabled Aggregator (RHEA) [22] is an expert-curated knowledgebase of chemical and transport reactions of biological interest - and the standard for enzyme and transporter annotation in UniProtKB. It provides a user-friendly interface to search, visualize and analyze the data. RHEA is freely accessible online.

ChEBI Chemical Entities of Biological Interest (ChEBI) [18] is a database containing information on small molecular entities, including their structures,

names, synonyms, and physicochemical properties, like their chemical form. It provides a systematic, manually curated classification of chemical entities based on their structural and chemical features. ChEBI is often used with other molecular biology databases, such as UniProt, to help annotate and analyze biological systems. Researchers in bioinformatics, systems biology, and drug discovery widely use it.

2.1.1 Choosing relevant data

Relevant data for a database of reactions catalyzed by CYPs for machine-learning applications include the name of the enzyme, its UniProt ID, information about reaction substrates and products, their ChEBI IDs, their SMILES[1], the reaction equation, and RHEA ID. Additionally, relevant data includes the sequence of the enzyme and any appropriate cofactors or reducing agents of the reaction. We also want to see in which kingdom these enzymes act, so we need to include the kingdom name specific for a specific enzyme and the organism name. In our results, we would also like to see the distribution of lengths of these enzymes, which means we also include the length of each sequence.

2.1.2 Data collection process

Once we identified relevant data for our database, we collected it from the above-mentioned databases.

For data collection, we used mainly the UniProt database as it already contains cross-references to the other two databases used and most of the chosen relevant data.

We downloaded a JSON file of all *reviewed*[2] cytochrome P450 enzymes from the UniProt database. We searched for the keyword - "PF00067", which is a Pfam domain specific for cytochrome P450 enzymes [25, 26] as mentioned in 1.2.1. We also tried searching for the cytochrome P450 superfamily, which gave us the same results as for the Pfam domain. To state the criteria for inclusion clearly:

- Only include reviewed entries (characterized) from the UniProt database
- To include only entries that are for sure cytochromes P450: search for the Pfam domain characteristic for these enzymes → PF00067

[1] SMILES (Simplified Molecular Input line Entry) is a chemical notation system designed to convert a description of a chemical structure to one unique notation. It allows rigorous structure notification by use of minimal and natural grammar [23, 24].

[2] a reviewed entry has been manually annotated and reviewed by UniProtKB curators, unreviewed entries are computer annotated

We extracted the primary accession number (the UniProt ID), protein name, cofactors and reaction details associated with each protein from the JSON file.

The extraction includes separating the reaction into substrates, products and reducing agents while removing unwanted participants, such as H(+), H2O and other unspecific participants of the specific enzyme. We know that the substrates and the reducing agents are on the left-hand side of the reaction equation, and the products are on the right. The extraction process also searches for cross-references to external databases, such as RHEA, and extracts the RHEA ID associated with the reaction. We stored all the data in a Pandas [27] DataFrame for further additions of relevant data. If no reaction or cofactor details are found for a protein in the JSON file, the data is still added to the data frame but missing important information about the reaction. We left it in the database for future completion with the use of different databases like BRENDA[3], KEGG[4], Atlas of Biochemistry[5], or BFD[6]. So far, not much relevant information has been found about reactions catalyzed by these enzymes.

Other information needed then includes substrate ChEBI ID, substrate SMILES, product ChEBI ID and product SMILES. We got this information from a file named reaction_data.csv, created by our lab colleague. It contains data from the RHEA database. In this file, each row represents one participant in the reaction. We had to think of how to access only the ones needed for our database. I created a dictionary with tuples (UniProt ID, RHEA ID) as a key and searched through reaction_data.csv for these tuples. In this case, the order of the reaction participant is the same as the order of the row with its ChEBI ID and SMILES containing the specific tuple. I first tried to get the ChEBI ID from the cross-references in UniProt, but the order was different.

The next step was adding the sequence, taxonomy information (organism name, kingdom name) and PFAM domains of each sequence. We used a lab colleague's script to download the taxonomy information (taxa.csv) and PFAM domains (pfams.csv) and adapted his script to download sequences (sequences.csv) from the UniProt database. After downloading, we added the information to our data frame. The final step was adding the sequence length to our data frame.

The source code containing the data collection can be found at GitHub[7]. [31].

[3]BRENDA provides reliable data, continuous curation and updates of classified enzymes, and the integration of newly discovered enzymes [28].

[4]KEGG is an integrated database resource for linking sequences to biological functions from molecular to higher levels [29].

[5]A database of all theoretical biochemical reactions based on known biochemical principles and compounds [30].

[6]Big Fantastic Database containing 2.5 billion sequences from UniProt/TrEMBL+Swissprot, Metaclust and Soil Reference Catalog Marine Eukaryotic Reference Catalog assembled by Plass

[7]https://github.com/komorn/BcProject.git

2.1.3 Manual addition of data

Some of the reactions we had in our database were the most general ones for CYPs. By the most general reaction, we mean the one mentioned in 1.4:

an organic molecule (RH) + O2 + reduced [NADPH–hemoprotein reductase]

=

an alcohol (R-OH) + H(+) + H2O + oxidized [NADPH–hemoprotein reductase].

We commonly do not want reactions like this in our database because we want the reactions to be specific. Additionally, we need them to contain all the needed information for machine-learning approaches, mainly the SMILES of the substrates and products.

We edited these reactions manually by adding a more specific substrate and product from the functional description in UniProt. It was only possible for about half of these reactions.

Other manual addition of data contained editing some of the kingdom's characteristics for individual enzymes in our database. The reason behind this was that from the taxa.csv file, we sometimes did not have a kingdom value. We searched for the taxonomy information at the NCBI's Taxonomy Browser.

2.2 Structure of the database

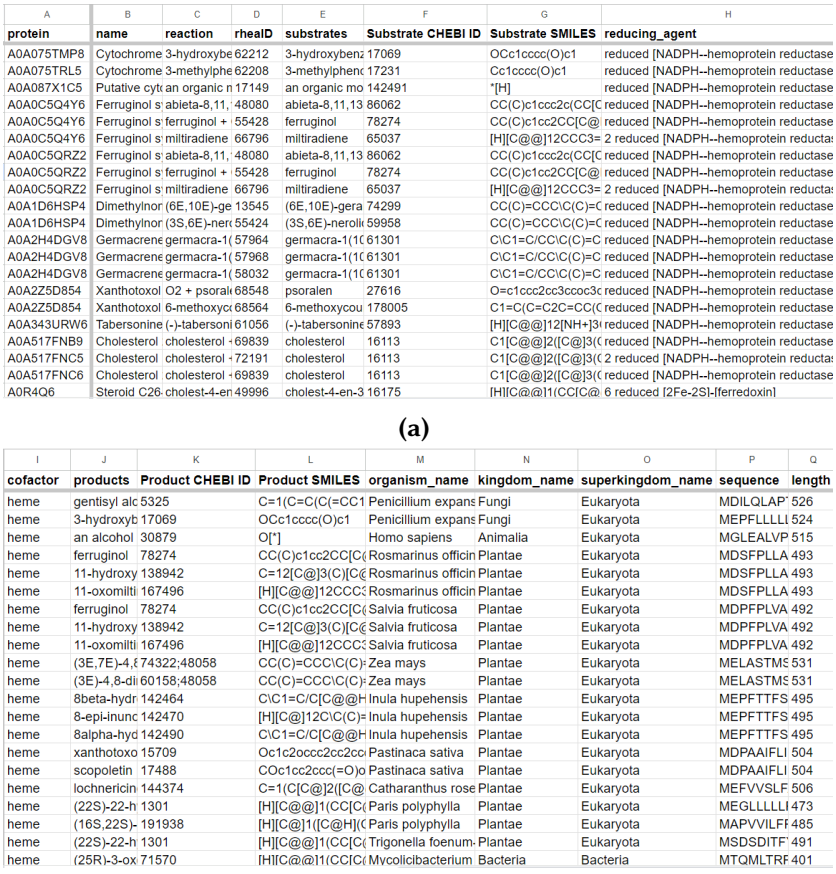
After taking the abovementioned steps, we finally have a database containing these columns:

- protein UniProt ID
- protein name
- reaction
- RHEA ID
- substrate
- substrate ChEBI ID
- substrate SMILES
- reducing agent
- cofactor
- product
- product ChEBI ID
- product SMILES
- organism name
- kingdom name
- superkingdom name
- PFAM
- sequence
- length

The database contains 2925 fully annotated reactions. 1726 unique enzymes catalyze these reactions.

Overall, we have 3780 records in the database. We kept the other less than 1000 records in the database for future manual addition of data from other databases like BRENDA, KEGG, Atlas of Biochemistry, BFD or from the function description in UniProt.

The database is, as of now, available as a .csv file on my previously mentioned GitHub and also as a .xlsx file on Google Drive[8].



A	B	C	D	E	F	G	H
protein	name	reaction	rheaID	substrates	Substrate CHEBI ID	Substrate SMILES	reducing_agent
A0A075TMP8	Cytochrome	3-hydroxybe	62212	3-hydroxybenz	17069	OCc1cccc(O)c1	reduced [NADPH--hemoprotein reductase
A0A075TRL5	Cytochrome	3-methylphe	62208	3-methylphenc	17231	Cc1cccc(O)c1	reduced [NADPH--hemoprotein reductase
A0A087X1C5	Putative cyt	an organic n	17149	an organic mo	142491	*[H]	reduced [NADPH--hemoprotein reductase
A0A0C5Q4Y6	Ferruginol s	abieta-8,11,	48080	abieta-8,11,13	86062	CC(C)c1ccc2c(CC[C	reduced [NADPH--hemoprotein reductase
A0A0C5Q4Y6	Ferruginol s	ferruginol +	55428	ferruginol	78274	CC(C)c1cc2CC[C@	reduced [NADPH--hemoprotein reductase
A0A0C5Q4Y6	Ferruginol s	miltiradiene	66796	miltiradiene	65037	[H][C@@]12CCC3=	2 reduced [NADPH--hemoprotein reducta
A0A0C5QRZ2	Ferruginol s	abieta-8,11,	48080	abieta-8,11,13	86062	CC(C)c1ccc2c(CC[C	reduced [NADPH--hemoprotein reductase
A0A0C5QRZ2	Ferruginol s	ferruginol +	55428	ferruginol	78274	CC(C)c1cc2CC[C@	reduced [NADPH--hemoprotein reductase
A0A0C5QRZ2	Ferruginol s	miltiradiene	66796	miltiradiene	65037	[H][C@@]12CCC3=	2 reduced [NADPH--hemoprotein reducta
A0A1D6HSP4	Dimethylnor	(6E,10E)-ge	13545	(6E,10E)-gera	74299	CC(C)=CCC\C(C)=C	reduced [NADPH--hemoprotein reductase
A0A1D6HSP4	Dimethylnor	(3S,6E)-ner	55424	(3S,6E)-neroli	59958	CC(C)=CCC\C(C)=C	reduced [NADPH--hemoprotein reductase
A0A2H4DGV8	Germacrene	germacra-1(	57964	germacra-1(1(	61301	C\C1=C/CC\C(C)=C	reduced [NADPH--hemoprotein reductase
A0A2H4DGV8	Germacrene	germacra-1(	57968	germacra-1(1(	61301	C\C1=C/CC\C(C)=C	reduced [NADPH--hemoprotein reductase
A0A2H4DGV8	Germacrene	germacra-1(	58032	germacra-1(1(	61301	C\C1=C/CC\C(C)=C	reduced [NADPH--hemoprotein reductase
A0A2Z5D854	Xanthotoxol	O2 + psoral	68548	psoralen	27616	O=c1ccc2cc3ccoc3c	reduced [NADPH--hemoprotein reductase
A0A2Z5D854	Xanthotoxol	6-methoxyc	68564	6-methoxycou	178005	C1=C(C=C2C=CC((	reduced [NADPH--hemoprotein reductase
A0A343URW6	Tabersonine	(-)-tabersoni	61056	(-)-tabersonine	57893	[H][C@@]12[NH+]3	reduced [NADPH--hemoprotein reductase
A0A517FNB9	Cholesterol	cholesterol +	69839	cholesterol	16113	C1[C@@]2([C@]3((	reduced [NADPH--hemoprotein reductase
A0A517FNC5	Cholesterol	cholesterol +	72191	cholesterol	16113	C1[C@@]2([C@]3((	2 reduced [NADPH--hemoprotein reducta
A0A517FNC6	Cholesterol	cholesterol +	69839	cholesterol	16113	C1[C@@]2([C@]3((	reduced [NADPH--hemoprotein reductase
A0R4Q6	Steroid C26	cholest-4-en	49996	cholest-4-en-3	16175	[H][C@@]1(CC[C@	6 reduced [2Fe-2S]-[ferredoxin]

(a)

I	J	K	L	M	N	O	P	Q
cofactor	products	Product CHEBI ID	Product SMILES	organism_name	kingdom_name	superkingdom_name	sequence	length
heme	gentisyl alc	5325	C=1(C=C(C(=CC1	Penicillium expans	Fungi	Eukaryota	MDILQLAP	526
heme	3-hydroxyb	17069	OCc1cccc(O)c1	Penicillium expans	Fungi	Eukaryota	MEPFLLLLI	524
heme	an alcohol	30879	O[*]	Homo sapiens	Animalia	Eukaryota	MGLEALVP	515
heme	ferruginol	78274	CC(C)c1cc2CC[C(	Rosmarinus officin	Plantae	Eukaryota	MDSFPLLA	493
heme	11-hydroxy	138942	C=12[C@]3(C)[C@	Rosmarinus officin	Plantae	Eukaryota	MDSFPLLA	493
heme	11-oxomiltii	167496	[H][C@@]12CCC3	Rosmarinus officin	Plantae	Eukaryota	MDSFPLLA	493
heme	ferruginol	78274	CC(C)c1cc2CC[C(	Salvia fruticosa	Plantae	Eukaryota	MDPFPLVA	492
heme	11-hydroxy	138942	C=12[C@]3(C)[C@	Salvia fruticosa	Plantae	Eukaryota	MDPFPLVA	492
heme	11-oxomiltii	167496	[H][C@@]12CCC3	Salvia fruticosa	Plantae	Eukaryota	MDPFPLVA	492
heme	(3E,7E)-4,8	74322;48058	CC(C)=CCC\C(C)=	Zea mays	Plantae	Eukaryota	MELASTMS	531
heme	(3E)-4,8-dii	60158;48058	CC(C)=CCC\C(C)=	Zea mays	Plantae	Eukaryota	MELASTMS	531
heme	8beta-hydr	142464	C\C1=C/C[C@@H	Inula hupehensis	Plantae	Eukaryota	MEPFTTFS	495
heme	8-epi-inunc	142470	[H][C@]12C\C(C)=	Inula hupehensis	Plantae	Eukaryota	MEPFTTFS	495
heme	8alpha-hyd	142490	C\C1=C/C[C@@H	Inula hupehensis	Plantae	Eukaryota	MEPFTTFS	495
heme	xanthotoxo	15709	Oc1c2occc2cc2cc	Pastinaca sativa	Plantae	Eukaryota	MDPAAIFLI	504
heme	scopoletin	17488	COc1cc2ccc(=O)o	Pastinaca sativa	Plantae	Eukaryota	MDPAAIFLI	504
heme	lochnericin	144374	C=1(C[C@]2([C@	Catharanthus rose	Plantae	Eukaryota	MEFVVSLF	506
heme	(22S)-22-h	1301	[H][C@@]1(CC[C(	Paris polyphylla	Plantae	Eukaryota	MEGLLLLLI	473
heme	(16S,22S)-	191938	[H][C@]1([C@H](	Paris polyphylla	Plantae	Eukaryota	MAPVVILFI	485
heme	(22S)-22-h	1301	[H][C@@]1(CC[C(	Trigonella foenum-	Plantae	Eukaryota	MSDSDITF	491
heme	(25R)-3-ox	71570	[H][C@@]1(CC[C(	Mycolicibacterium	Bacteria	Bacteria	MTQMLTRF	401

(b)

Figure 2.1 Screenshot of the final database for illustration.

[8]Google Drive link

Chapter 3

Analysis of the database

This chapter presents an analysis of our created database, which includes information on the sequences, substrates and kingdoms. We investigate the distribution of CYP sequences across different kingdoms and the length distribution of the sequences. Furthermore, we compare the results with those of all characterized and uncharacterized CYPs from the UniProt database[1].

How did we download all CYP sequences from UniProt

To download all CYP sequences, we used a Python library called Bioinformatic-data-collector[2]. Our colleague created this library. This tool is used for the generation and execution of SPARQL[3] queries for usage in bioinformatics. The script we used to download this data is available on my previously mentioned GitHub. It is the get_data.py script, also created by our colleague. It defines a function collect_data which uses the requests module to send a query to a specified URL and retrieve the results as a Pandas DataFrame. It defines a function get_uniprot which retrieves data on UniProt proteins using the UniprotQueryBuilder and UniprotSearchConfig classes from the bioinfdatacollector module. This function uses a filter on PF00067 to include only CYPs. It defines a function get_query which constructs a SPARQL query. In the __main__ block, the script calls the function get_uniprot three times to retrieve taxa, protein names, and protein domains (PFAMs) data. I added a fourth call of this function to retrieve the sequences. It then retrieves reaction data using the get_uniprot function and the ReactionQueryBuilder and ReactionSearchConfig classes from the bioinfdatacollector module. The script then splits the list of reaction IDs into

[1] contains 484323 entries

[2] Bioinformatic-data-collector (bioinfdatacollector module)

[3] the standard query language and protocol for Linked Open Data on the web or for RDF triplestores.

batches of 20, constructs Rhea search queries for each batch, and retrieves the corresponding data using the collect_data function. The script then joins the UniProt and Rhea data frames on the reaction ID and saves the resulting data frame as a CSV file to reaction_data.csv. This data contains both reviewed and unreviewed entries for PF00067, unlike our database, which only contains reviewed entries (we refer to reviewed as characterized and unreviewed as uncharacterized).

We also check the correctness of the results from all characterized and uncharacterized CYPs with data in the InterPro database. We are only comparing the data from these two databases for the number of CYP sequences in different kingdoms. The reason behind this is to compare the data obtained from the UniProt database by searching for a Pfam domain with the data available in a database of protein families for the same Pfam domain.

Finally, we classify the substrates into different chemical classes and present a visualization of the chemical space of the substrates. Our analysis provides insights into the diversity and substrate specificity of CYPs and the comparison between the characterized and uncharacterized CYPs.

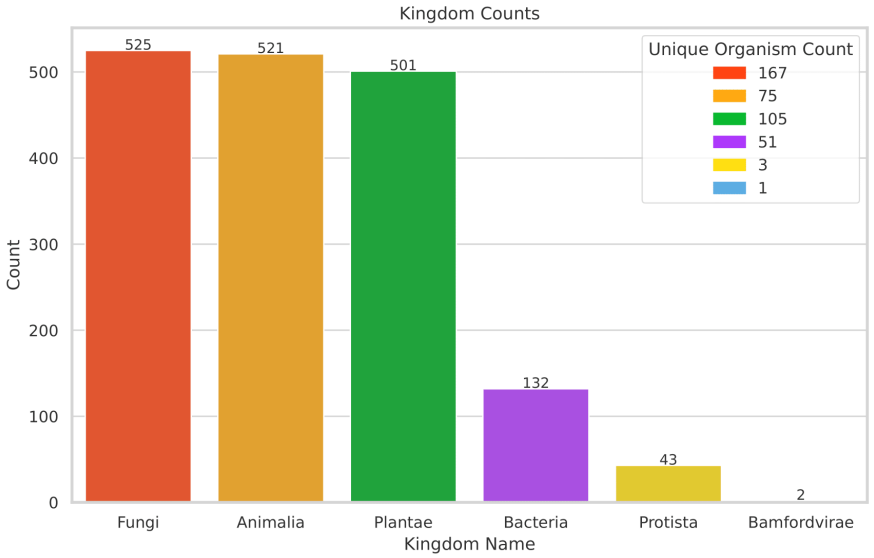
3.1 Number of sequences in kingdoms

Figure 3.1 shows the number of CYP sequences in different kingdoms. We can see that the kingdom with the most sequences in our and the UniProt databases is Fungi.

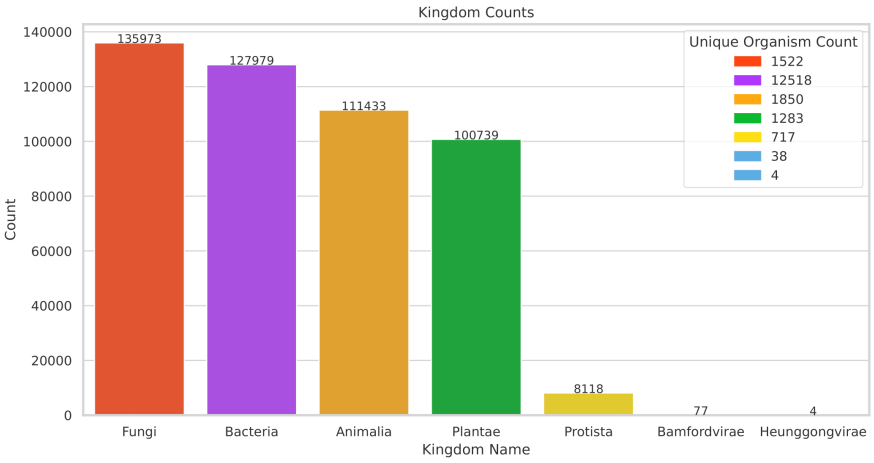
The most significant difference we can see between figures 3.2a and 3.2b is in Bacteria, where we only have 132 sequences in our database of characterized CYPs. However, it is the kingdom with the second-highest number of sequences in the UniProt database with both characterized and uncharacterized CYPs, which means that the most uncharacterized and unreviewed CYP sequences are believed to be from the Bacteria kingdom.

The idea that viral genomes could contain CYP genes was not considered until the discovery of giant viruses revealed multiple and distinct CYP genes [32]. That is the reason for the lowest number of sequences from Viruses (Bamfordvirae and Heunggongvirae).

A surprising result is that Animalia have more sequences than Plantae. From the numbers we provide for the number of CYP genes in different organisms in the next section 3.2, it should be the other way around. The reason behind this is that there are many more sequenced genomes in animals than in plants. As many as 3,278 unique animals have had their nuclear genome sequenced [33], while only 1,139 genomes from green plants have been sequenced [34]. Here, for our database of characterized CYPs, the average number of sequences per organism



(a) Characterized CYPs database



(b) Characterized and uncharacterized CYPs

Figure 3.1 The number of CYP sequences in different kingdoms. The label in the upper right of the figures shows how many unique organisms there are in each kingdom for the given number of sequences.

from the Animalia kingdom is almost 7. The average number of sequences per organism from the Plantae kingdom is approximately 4.8. Figures 3.1a and 3.1b show the number of unique organisms per all sequences in each kingdom.

Comparison between the characterized and uncharacterized CYPs we obtained from the UniProt database and the data from the InterPro database can be seen in Table 3.1. The numbers of sequences for Protista, Bamfordvirae, and Heunggongvirae are included in the Others category in the InterPro database, so we can not provide the actual number of sequences here.

Kingdom	UniProt data	InterPro data
Fungi	135973	122518
Bacteria	127979	130149
Animalia	111433	110034
Plantae	100739	98518
Protista	8118	no data
Bamfordvirae	77	no data
Heunggongvirae	4	no data

Table 3.1 Comparison between UniProt data for characterized and uncharacterized CYPs vs. InterPro data.

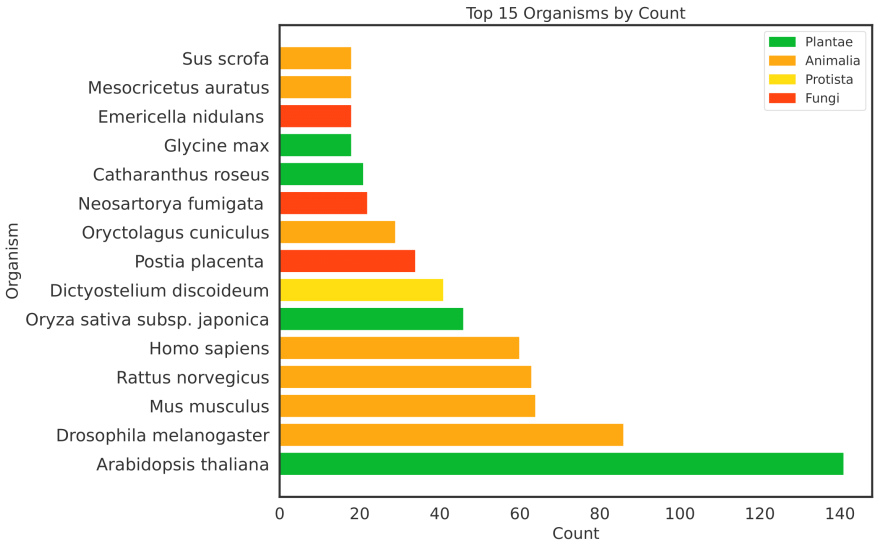
The main question is why there is such a big difference between the number of sequences in our data and the InterPro data in Fungi. This may be caused by the different database we used (the UniProt), but we still do not think it should make such a big difference.

3.2 Organisms with the most CYP sequences

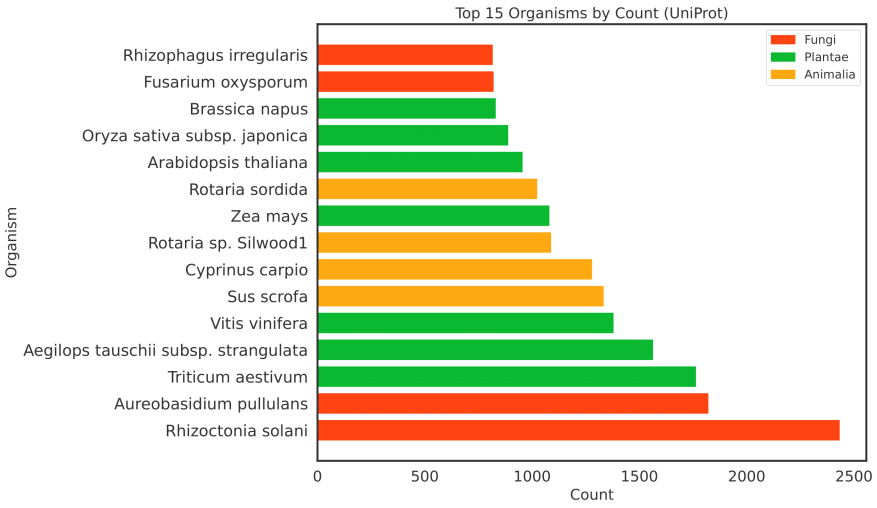
Figure 3.2 shows which species from our and UniProt databases have the most CYP sequences.

In the first subplot 3.2a, we can see that the most abundant organism is *Arabidopsis thaliana*, which is known to have 244 CYP genes [35]. In the second subplot 3.2b it is the fifth organism from plants with the most sequences. Some plants are known to have more CYP genes than *Arabidopsis*. These organisms include *Triticum aestivum* (more than 1000 isoforms are being expected [36]) or *Vitis vinifera* (579 CYP genes [37]). Both can be seen in Figure 3.2b.

Humans have 57 CYP genes [38] - corresponding to the number of sequences in Figure 3.2a from our database of characterized CYPs. We cannot see human (*Homo sapiens*) in the subplot of all characterized and uncharacterized CYPs.



(a) Characterized CYPs database



(b) Characterized and uncharacterized CYPs

Figure 3.2 Top 15 organisms by the count of CYP sequences.

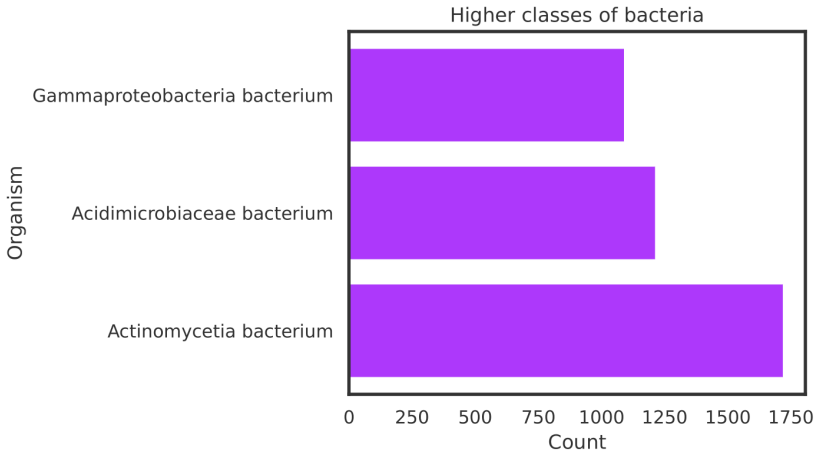


Figure 3.3 Higher bacterial classes that are very common in CYPs.

Drosophilia melanogaster has 87 CYP genes [39], which also corresponds to the number of sequences obtained from our database.

Animals dominate in the first subplot 3.2a from our database, whereas plants dominate in the second one 3.2b with nine organisms from the Plantae kingdom compared to only four organisms that belong to the Animalia kingdom.

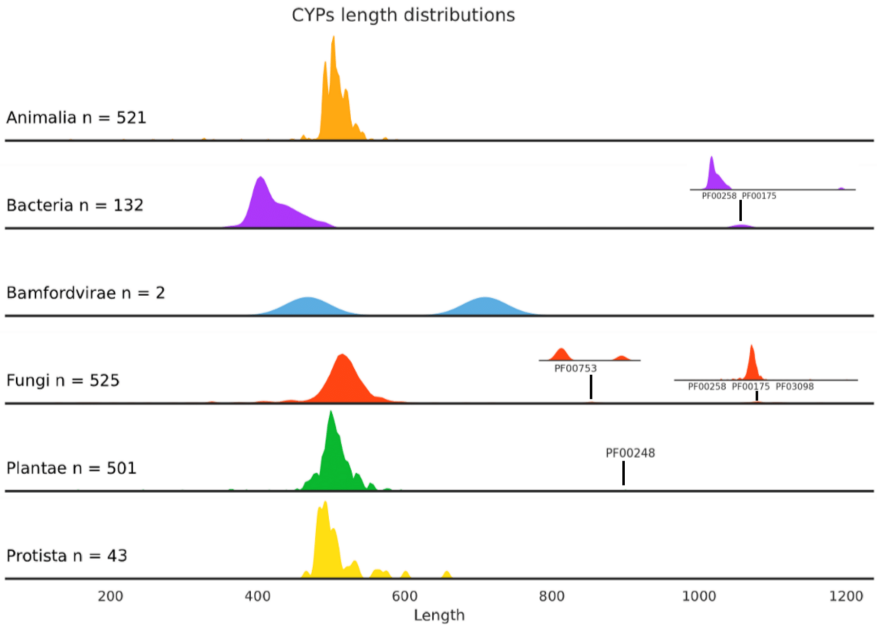
In the second subplot 3.2b, we can see that the two organisms with the most CYP sequences are Fungi.

We also show three higher classes of bacteria in Figure 3.3 with as many CYP sequences as some of the organisms shown in Figure 3.2. We put these in a separate figure because they are not species but classes or families.

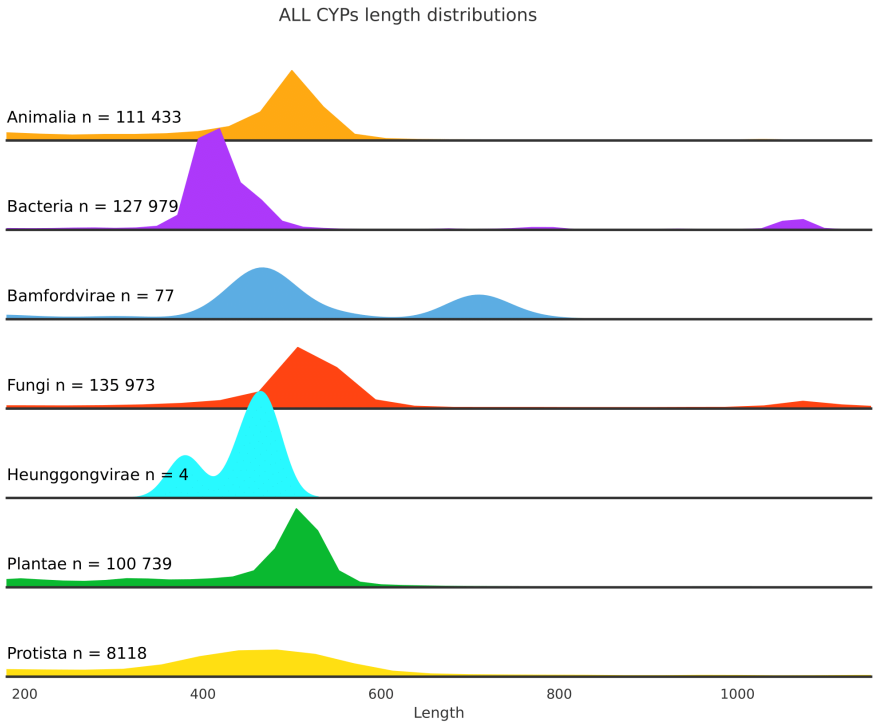
3.3 The distribution of the length of amino acid sequences

The analysis of the distribution of CYP sequence length in different kingdoms is crucial because it can provide insights into the functional diversity of these enzymes. We can see the length distributions in Figure 3.4. The first subfigure 3.4a shows the length distributions of characterized CYPs. The most characterized sequences are about 500 amino acids long. We can see an exception in the Bacteria kingdom where sequences are shorter. This may be caused by incomplete or fragmented sequences. There are a lot of bacterial genomes and often it is just incomplete data, or metagenomic sequencing[4]. This is also probably causing the

[4]the application of sequencing techniques to analyse the totality of the genomic material present in a sample [40]



(a) Characterized CYPs database



(b) Characterized and uncharacterized CYPs

Figure 3.4 Length distribution of CYPs amino acid sequences in different kingdoms. The letter n by the kingdom labels indicates the number of sequences.

high number of bacterial and fungi sequences in 3.1.

Small peaks with longer enzymes are observed around 800 amino acids long in Fungi and Plantae. In Plantae, with only one enzyme, a bifunctional protein with two modules (cytochrome P450, oxidoreductase module) has 901 amino acids. This enzyme has an additional PFAM domain, the PF00248. This PFAM domain corresponds to the Aldo/keto reductase family. In Fungi, this peak corresponds to PF00753.

Fungi and Bacteria have a small peak of around 1000 amino acids, which can be attributed to some bifunctional proteins or parts of gene clusters. In Fungi, these sequences correspond to PFAMs PF00258, PF00175, and PF03098, while in bacteria, they correspond to PF00258 and PF00175. Some of these enzymes from the Fungi kingdom are hybrid enzymes, with two enzymes connected in one sequence. Examples of these enzymes are Q9HGE0, A0A3G9HRC2, and A0A455R5H4. All of these contain the PF00175 domain.

The mentioned PFAM domains are:

- PF00258 - Flavodoxin/nitric oxide synthase domain
- PF00175 - Oxidoreductase FAD/NAD(P)-binding domain
- PF03098 - Animal haem peroxidase domain
- PF0753 - Metallo-beta-lactamase
- PF00248 - Aldo/keto reductase family

Overall, we found that the long sequences contain two PFAM domains, the main one and the ones shown in Figure 3.4a. These observations suggest that the length and domain composition of CYPs can vary greatly depending on the organism and the specific biological functions required.

For Figure 3.4b, we can see the same pattern of peaks as in the first figure 3.4a created from CYPs from our database. The only difference is that we can see more short sequences, which are probably just some fragments of a sequence, not a complete sequence.

For further analysis, we only used the created database of characterized CYPs.

3.4 Substrate classification

Classifying substrates from our database allows a better understanding of the chemical space these enzymes can accommodate and modify. Substrate classification can help to identify trends and patterns in substrate specificity across

different cytochrome P450 enzymes, which can lead to new insights into the function of these enzymes and also into predicting their function.

For the classification of substrates from our database, we used the NPClassifier [41], a deep neural network for classifying natural products.

The results can be seen in Figure 3.5. We also show how these classes are distributed into different kingdoms. The first subfigure 3.5a offers all the present classes of substrates in our database. We can see that Terpenoids and Steroids is the class with the highest count, where steroids mainly comprise the Animalia kingdom's substrates and terpenoids comprise part of the Plantae kingdom. In animals, terpenoid precursors produce steroids and sterols [42]. Terpenoids are the largest class of plant secondary metabolites, representing about 60% of known natural products [43]. The second most common class is Fatty acids. Cytochrome P450 enzymes catalyze the hydroxylation of fatty acids, and this metabolic pathway is mainly present in animals. Our classification results show that most of the fatty acid substrates are from the Animalia kingdom.

The distribution of the other substrate classes in kingdoms can't be adequately seen from the first subfigure. We provide a close-up of the substrate classes with lower counts in subfigure 3.5b.

While in the first two substrate classes, we could have seen Animals dominating, Figure 3.5b shows that other kingdoms dominate. Plants are the most abundant in the Shikimates, Phenylpropanoids, and Alkaloids classes. This is a direct result of the facts that the shikimate pathway is not present in animals [44], the phenylpropanoid pathway [45] produces mainly compounds involved in plant defence, and the variety of structures of alkaloids is the highest in plants [46].

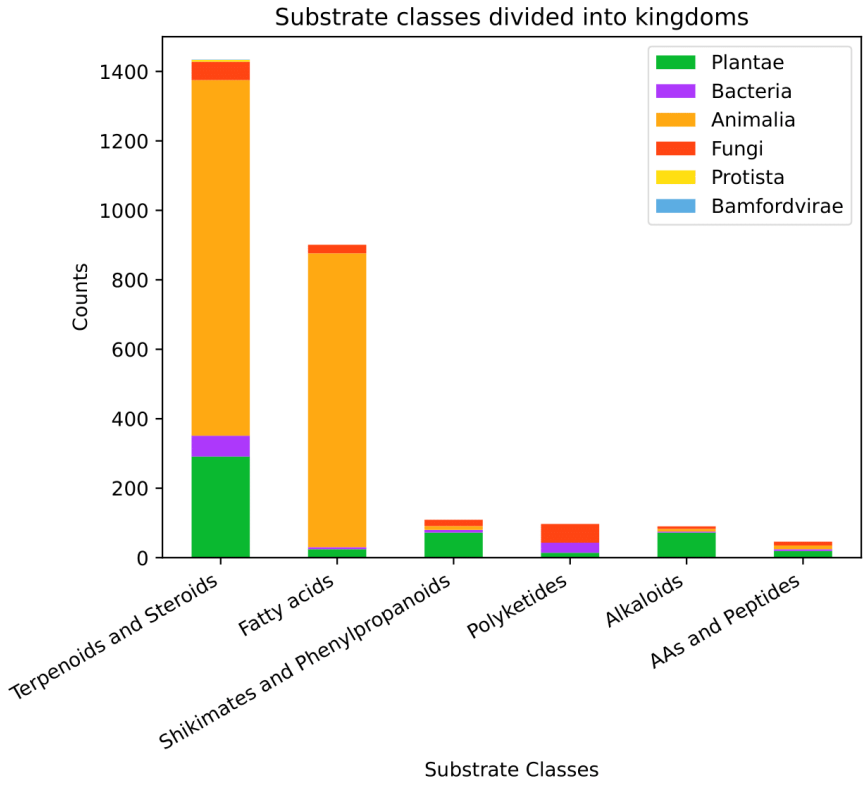
3.5 Visualization of the chemical space of the substrates

The last analysis we have done for this thesis is the visualization of the chemical space of the substrates.

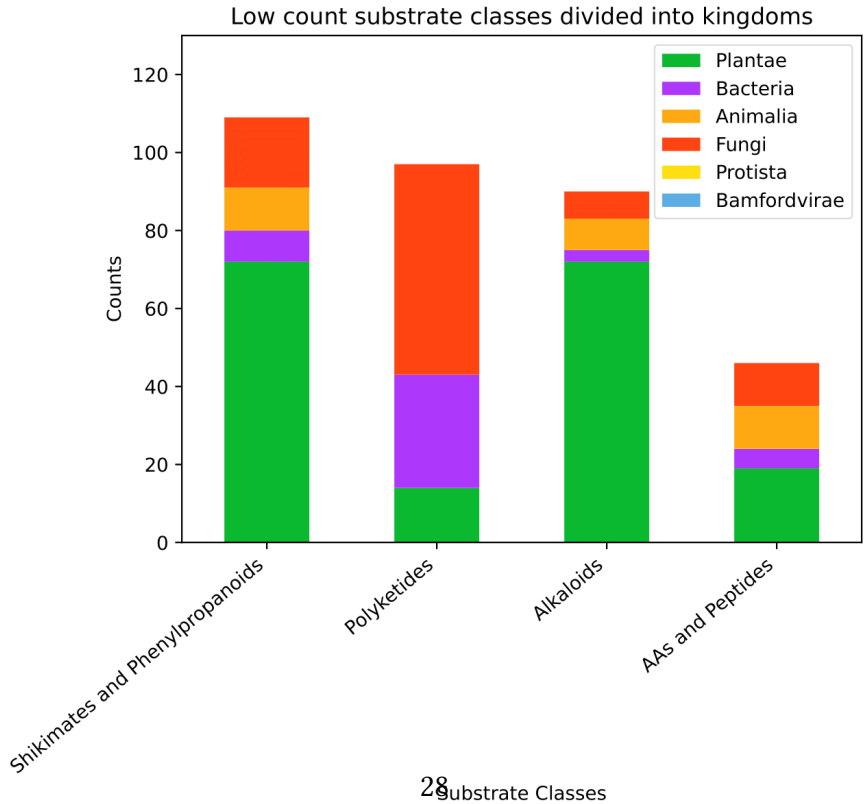
By using a TMAP algorithm[5], we generated a hierarchical clustering tree, representing the relationships between the substrates in our database in a high-dimensional space. The visualization can be seen in Figure 3.6.

The main reason for creating this visualization was to see if the plant CYPs, with higher specificity for substrates, take up more of the chemical space than the animal CYPs, which are more promiscuous.

[5]TMAP (Tree MAP) is an algorithm that generates and distributes intuitive visualizations of large data sets in the order of up to 10^7 with arbitrary dimensionality in a tree [47]



(a) All substrate classes



(b) Close up to substrate classes with lower counts

Figure 3.5 Substrate classes divided into kingdoms.

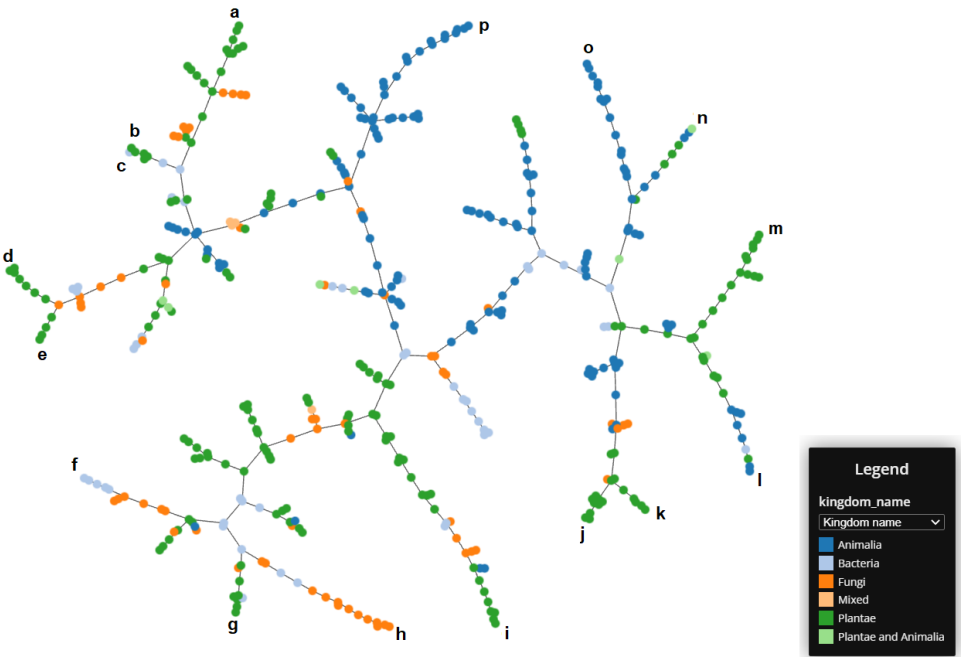
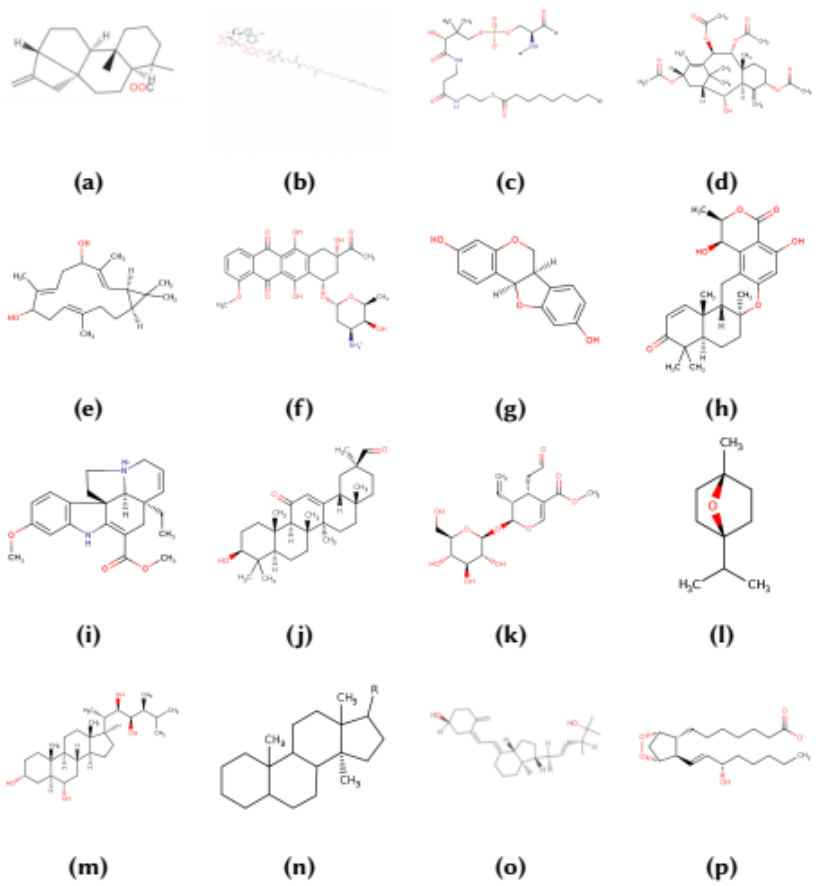


Figure 3.6 Visualization of the redundancy of the substrates in different kingdoms. Each node represents a unique substrate from our database. The further these nodes are from each other, the larger the difference between these chemical structures. Legend shows the colors of different kingdoms.



3.7a - ent-kaur-16-en-19-oate: a monocarboxylic acid anion (kaurene diterpenoid); 3.7b - unknown; 3.7c - a C2-C8-saturated long-chain fattyacyl-[ACP]: a fatty acid; 3.7d - 2αhydroxytaxusin: a taxane diterpenoid; 3.7e - 4,8-dihydroxycasbene; 3.7f - daunorubicin(1+): an anthracycline antibiotic; 3.7g - (6aR,11aR)-3,9-dihydroxypterocarpan: a pterocarpan, which is a derivative of isoflavonoids found in the family Fabaceae; 3.7h - chrodrimanin T: a heteropentacyclic compound; 3.7i - 16-methoxytabersoninium(1+): a monoterpenoid indole alkaloid; 3.7j - glycyrrhetaldehyde: a triterpenoid; 3.7k - (-)-secologanin: an iridoid monoterpenoid; 3.7l - 1,4-cineole: an oxabicycloalkane; 3.7m - 6α-hydroxytyphasterol: a brassinosteroid; 3.7n - 14α-methyl steroid: a steroid carrying a 14α-methyl substituent; 3.7o - unknown; 3.7p - prostaglandin: an eicosanoid having diverse hormone-like effects in animals. Figure adapted from [18]

Figure 3.7

We again used different colours for each kingdom to see how much chemical space the substrates from various kingdoms take up. When we look closer at the tree, we can see far more nodes from the Animalia and Plantae kingdoms than the others. We can also see green (Plants) in more branches of the tree than blue (Animals).

The number of unique substrates from our database is 456. The number of substrates from Plants is 179, and 149 from Animals. So there is a slight difference between Plants and Animals.

In Figure 3.6, the end branches are marked with a letter, indicating their designation in the following Figure 3.7. We created this labelling and figure to understand better what substrates are present. Most of the marked substrates are terpenoids, which corresponds to the results we got from our substrate classification.

However, the overall result of the visualization is not what we expected. We expected much more substrates in the Plantae kingdom than in the Animalia kingdom. We also assumed that the tree would be more spread out. However, this is a direct consequence of our database containing only 456 unique substrates.

3.6 Future

Our future concerns will begin by continuing with the manual addition of data on reactions catalyzed by the enzymes we have in our database but currently do not have information about the substrates and products. This information will be added from other sources mentioned in 2.1.2. We will then continue by adding the publication links to the characterized reactions in our database. After these steps, we also want to add the type of reaction to every reaction in our database. We plan to expand a developed machine learning-based approach for predicting an enzyme's function to cytochromes P450. However, there is a need to develop a better method that will be able to recognize enzymes that are highly promiscuous, like CYPs, which can accommodate thousands of substrates.

The next step will be sharing the database with other research groups to provide a comprehensive database of reactions catalyzed by these crucial enzymes in all living organisms.

Conclusion

This thesis provides a comprehensive overview of cytochrome P450 enzymes, including their nomenclature, architecture, functional domains, and promiscuity. We discussed the catalytic cycle of P450 enzymes and some of the basic and plant-specific reactions they catalyze.

We have presented the process of developing a database of reactions catalyzed by cytochrome P450 enzymes in chapter 2. The database, as of now, contains 2925 fully annotated reactions catalyzed by 1726 unique CYPs. This is a valuable source for our planned machine-learning applications, mainly involving predicting the function of the uncharacterized CYPs. The database will be further updated as new reactions catalyzed by CYPs are characterized.

Furthermore, in chapter 3, we analyzed the developed database regarding distribution among the kingdoms of organisms, distribution of lengths of amino acid sequences, substrate classification, and chemical space visualization. The results we obtained by analyzing the database were surprising, especially in how many characterized CYPs there are in animals and how many in plants. However, this is a direct consequence of having far fewer plant genomes than animal genomes sequenced. Generally, plants tend to have more CYPs than animals, as plants have evolved to produce various secondary metabolites, many of which require CYPs for their biosynthesis. Additionally, plants need detoxification enzymes more due to their sedentary lifestyle.

Bibliography

[1] David R Nelson. "The cytochrome p450 homepage". In: *Human genomics* 4.1 (2009), pp. 1–7.

[2] Zhong Li et al. "Engineering cytochrome P450 enzyme systems for biomedical and biotechnological applications". In: *Journal of Biological Chemistry* 295.3 (2020), pp. 833–849.

[3] Ohgew Kweon et al. "CYPminer: an automated cytochrome P450 identification, classification, and data analysis tool for genome data sets across kingdoms". In: *BMC bioinformatics* 21.1 (2020), pp. 1–11.

[4] Paul R Ortiz De Montellano et al. *Cytochrome P450: structure, mechanism, and biochemistry*. Vol. 3. Springer, 2005.

[5] LE Khmelevtsova et al. "Prokaryotic cytochromes P450". In: *Applied Biochemistry and Microbiology* 53 (2017), pp. 401–409.

[6] DJ Cook et al. "Cytochromes P450: history, classes, catalytic mechanism, and industrial application". In: *Advances in protein chemistry and structural biology* 105 (2016), pp. 105–126.

[7] W. Nam. "8.12 - Cytochrome P450". In: *Comprehensive Coordination Chemistry II*. Ed. by Jon A. McCleverty and Thomas J. Meyer. Oxford: Pergamon, 2003, pp. 281–307. ISBN: 978-0-08-043748-4. DOI: https://doi.org/10.1016/B0-08-043748-6/08145-7. URL: https://www.sciencedirect.com/science/article/pii/B0080437486081457.

[8] David R Nelson. "Cytochrome P450 diversity in the tree of life". In: *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* 1866.1 (2018), pp. 141–154.

[9] Typhaine Paysan-Lafosse et al. "InterPro in 2022". In: *Nucleic Acids Research* 51.D1 (2023), pp. D418–D427.

[10] Danièle Werck-Reichhart. "Promiscuity, a Driver of Plant Cytochrome P450 Evolution?" In: *Biomolecules* 13.2 (2023), p. 394.

[11] David FV Lewis and John M Pratt. “The P450 catalytic cycle and oxygenation mechanism”. In: *Drug metabolism reviews* 30.4 (1998), pp. 739–786.

[12] Tom Coleman et al. “Understanding the mechanistic requirements for efficient and stereoselective alkene epoxidation by a cytochrome P450 enzyme”. In: *ACS Catalysis* 11.4 (2021), pp. 1995–2010.

[13] Masaharu Mizutani and Fumihiko Sato. “Unusual P450 reactions in plant secondary metabolism”. In: *Archives of Biochemistry and Biophysics* 507.1 (2011), pp. 194–203.

[14] Balazs Siminszky. “Plant cytochrome P450-mediated herbicide metabolism”. In: *Phytochemistry Reviews* 5.2-3 (2006), pp. 445–458.

[15] Emre M Isin and F Peter Guengerich. “Complex reactions catalyzed by cytochrome P450 enzymes”. In: *Biochimica et Biophysica Acta (BBA)-General Subjects* 1770.3 (2007), pp. 314–329.

[16] F Peter Guengerich. “Common and uncommon cytochrome P450 reactions related to metabolism and chemical toxicity”. In: *Chemical research in toxicology* 14.6 (2001), pp. 611–650.

[17] Bernard Meunier, Samuel P De Visser, and Sason Shaik. “Mechanism of oxidation reactions catalyzed by cytochrome P450 enzymes”. In: *Chemical reviews* 104.9 (2004), pp. 3947–3980.

[18] Janna Hastings et al. “ChEBI in 2016: Improved services and an expanding collection of metabolites”. In: *Nucleic acids research* 44.D1 (2016), pp. D1214–D1219.

[19] Eiichiro Ono et al. “Formation of two methylenedioxy bridges by a Sesamum CYP81Q protein yielding a furofuran lignan,(+)-sesamin”. In: *Proceedings of the National Academy of Sciences* 103.26 (2006), pp. 10116–10121.

[20] Rita Bernhardt and Vlada B Urlacher. “Cytochromes P450 as promising catalysts for biotechnological application: chances and limitations”. In: *Applied microbiology and biotechnology* 98 (2014), pp. 6185–6203.

[21] “UniProt: the Universal Protein knowledgebase in 2023”. In: *Nucleic Acids Research* 51.D1 (2023), pp. D523–D531.

[22] Parit Bansal et al. “Rhea, the reaction knowledgebase in 2022”. In: *Nucleic acids research* 50.D1 (2022), pp. D693–D700.

[23] David Weininger. “SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules”. In: *Journal of chemical information and computer sciences* 28.1 (1988), pp. 31–36.

[24] David Weininger, Arthur Weininger, and Joseph L Weininger. "SMILES. 2. Algorithm for generation of unique SMILES notation". In: *Journal of chemical information and computer sciences* 29.2 (1989), pp. 97–101.

[25] René De Mot and Annabel HA Parret. "A novel class of self-sufficient cytochrome P450 monooxygenases in prokaryotes". In: *Trends in microbiology* 10.11 (2002), pp. 502–508.

[26] Yang Liu et al. "Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes". In: *Genome biology* 5.9 (2004), pp. 1–11.

[27] Jeff Reback et al. "pandas-dev/pandas: Pandas 1.0. 5". In: *Zenodo* (2020).

[28] Antje Chang et al. "BRENDA, the ELIXIR core data resource in 2021: new developments and updates". In: *Nucleic acids research* 49.D1 (2021), pp. D498–D508.

[29] Minoru Kanehisa. "Enzyme annotation and metabolic reconstruction using KEGG". In: *Protein function prediction: methods and protocols* (2017), pp. 135–145.

[30] Noushin Hadadi et al. "ATLAS of biochemistry: a repository of all possible biochemical reactions for synthetic biology and metabolic engineering studies". In: *ACS synthetic biology* 5.10 (2016), pp. 1155–1166.

[31] Conrad L Schoch et al. "NCBI Taxonomy: a comprehensive update on curation, resources and tools". In: *Database* 2020 (2020).

[32] David C Lamb et al. "On the occurrence of cytochrome P450 in viruses". In: *Proceedings of the National Academy of Sciences* 116.25 (2019), pp. 12343–12352.

[33] Scott Hotaling, Joanna L Kelley, and Paul B Frandsen. "Toward a genome sequence for every animal: Where are we now?" In: *Proceedings of the National Academy of Sciences* 118.52 (2021), e2109019118.

[34] W John Kress et al. "Green plant genomes: What we know in an era of rapidly expanding opportunities". In: *Proceedings of the National Academy of Sciences* 119.4 (2022), e2115640118.

[35] Søren Bak et al. "Cytochromes P450". In: *The Arabidopsis Book/American Society of Plant Biologists* 9 (2011).

[36] Rita Bernhardt. "Cytochromes P450 as versatile biocatalysts". In: *Journal of biotechnology* 124.1 (2006), pp. 128–145.

[37] Daniela Minerdi, Stefania Savoi, and Paolo Sabbatini. “Role of Cytochrome P450 Enzyme in Plant Microorganisms’ Communication: A Focus on Grapevine”. In: *International Journal of Molecular Sciences* 24.5 (2023), p. 4695.

[38] Daniel W Nebert, Kjell Wikvall, and Walter L Miller. “Human cytochromes P450 in health and disease”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 368.1612 (2013), p. 20120431.

[39] Shane R Baldwin et al. “Identification and characterization of CYPs induced in the Drosophila antenna by exposure to a plant odorant”. In: *Scientific Reports* 11.1 (2021), p. 20530.

[40] Ana Elena Pérez-Cobas, Laura Gomez-Valero, and Carmen Buchrieser. “Metagenomic approaches in microbial ecology: an update on whole-genome and marker gene sequencing analyses”. In: *Microbial genomics* 6.8 (2020).

[41] Hyun Woo Kim et al. “NPClassifier: A deep neural network-based structural classification tool for natural products”. In: *Journal of Natural Products* 84.11 (2021), pp. 2795–2807.

[42] Beverly AS Reyes et al. “Selected phyto and marine bioactive compounds: Alternatives for the treatment of type 2 diabetes”. In: *Studies in Natural Products Chemistry* 55 (2018), pp. 111–143.

[43] Björn Hamberger and Søren Bak. “Plant P450s as versatile drivers for evolution of species-specific chemical diversity”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 368.1612 (2013), p. 20120426.

[44] Teresa Rocha-Santos and Armando C Duarte. “Introduction to the analysis of bioactive compounds in marine samples”. In: *Comprehensive Analytical Chemistry*. Vol. 65. Elsevier, 2014, pp. 1–13.

[45] Daniele Werck-Reichhart. “Cytochromes P450 in phenylpropanoid metabolism”. In: *Drug metabolism and drug interactions* 12.3-4 (1995), pp. 221–244.

[46] Jean Claude Braekman, Désiré Daloze, and JM Pasteels. “Alkaloids in animals”. In: *Alkaloids: Biochemistry, Ecology, and Medicinal Applications* (1998), pp. 349–378.

[47] Daniel Probst and Jean-Louis Reymond. “Visualization of very large high-dimensional data sets as minimum spanning trees”. In: *Journal of Cheminformatics* 12.1 (2020), pp. 1–13.

Appendix A

Github and Google Drive

The GitHub repository for this project contains several important files. The "allCYPS.csv" file includes all the characterized and uncharacterized entries from UniProt. The "ALLcyps.ipynb" file documents the process of merging data from "namesNAT.csv", "taxaNAT.csv", and "sequencesNAT.csv" to create "allCYPs.csv" file. The "cyps_analysis.ipynb" file analyzes data from our database as well as UniProt, generating the plots presented in chapter 3 of the thesis. Other important files generated by the "get_data.py" script include "names(NAT).csv", "taxa(NAT).csv", "sequences(NAT).csv", "pfams.csv", and "reaction_data.csv". The "npClassifier.ipynb" file is used for substrate classification. The "uniprot_json.json" file includes all the reviewed entries for PF00067. Python notebook "reaction-parsing.ipynb" extracts data from a downloaded JSON file and merges it with the data from the aforementioned ".csv" files to create the "reactionsfromjson.csv" file, which is manually edited to create the final database. The final database file is named "cypsDB-new.csv". Additionally, the "substrates.tsv" file contains all the substrates' SMILES, which is used for substrate classification. The "tmap.ipynb" file generates the chemical space visualization. There is also a folder with all generated plots for this thesis named img. A folder with the HTML file containing the TMAP tree is in the tmap folder.

A Google Sheets version of the final database is available on Google Drive[1]. This Google Drive also shows the plots used in this thesis. QR code below.

[1]Google drive with the final database