Charles University

Faculty of Science

Study programme: Bioinformatics

Branch of study: Bioinformatics

Bc. Tereza Čalounová

Mining novel terpene synthases from large-scale repositories

Mining nových terpen syntáz z rozsáhlých databází

Master thesis

Supervisor: Mgr. Tomáš Pluskal, Ph.D.

Prague, 2024

Prohlášení:

Prohlašuji, že jsem závěrečnou práci zpracovala samostatně a že jsem uvedla všechny použité informační zdroje a literaturu. Tato práce ani její podstatná část nebyla předložena k získání jiného nebo stejného akademického titulu.

V Praze, 30. 4. 2024 Tereza Čalounová

Acknowledgment

I would like to thank my supervisor, Tomáš Pluskal, for his guidance throughout this project and previous projects during my stay in his research group at IOCB Prague. I am grateful for the knowledge and experience gained. Additionally, I would like to thank all members of the Pluskal lab, as I had a great time working with them.

I would like to thank my family, friends, and especially Bernhard for their big support.

Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

Abstract

Terpenes and terpenoids represent the largest and most structurally diverse group of natural products, with applications across many fields, including the pharmaceutical industry. These molecules are synthesized in nature by enzymes known as terpene synthases. This thesis conducted a bioinformatic analysis of a curated database containing all 1125 experimentally characterized terpene synthases, focusing on identifying patterns in sequence lengths and domain architectures of these enzymes across different kingdoms of life.

Based on this analysis's knowledge, sequence-guided mining was conducted to identify possible new terpene synthases. Using nearly 5.5 billion protein sequences from various large-scale sequence repositories, the mining resulted in the identification of more than 600 thousand putative terpene synthases. These putative terpene synthases mainly originate from Bacteria and metagenomes, sources that had historically been less explored.

The resulting dataset, accompanied by a phylogenetic tree, sequence similarity network, and two prioritization scores, offers a valuable resource for the discovery of novel terpenes.

Keywords: terpene synthase, TPS, mining, Pfam, SUPERFAMILY, domain, terpene

Abstrakt

Terpeny a terpenoidy představují největší a strukturně nejrozmanitější skupinu přírodních látek s využitím v mnoha oborech, včetně farmaceutického průmyslu. Tyto molekuly jsou v přírodě syntetizovány enzymy známými jako terpen syntázy. V této práci byla provedena bioinformatická analýza kurátorované databáze obsahující všech 1125 experimentálně charakterizovaných terpen syntáz se zaměřením na identifikaci vzorců v délkách sekvencí a doménových architekturách těchto enzymů napříč různými říšemi života.

Na základě poznatků této analýzy byl proveden sekvenčně založený mining s cílem identifikovat možné nové terpen syntázy. S využitím téměř 5,5 miliard proteinových sekvencí z různých rozsáhlých sekvenčních databází vedl mining k identifikaci více než 600 tisíc potenciálních terpen syntáz. Tyto potenciální terpen syntázy pocházejí převážně z bakterií a metagenomů, tedy ze zdrojů, které byly historicky méně zkoumány.

Výsledný dataset, doplněný fylogenetickým stromem, sítí sekvenční podobnosti a dvěma skóre prioritizace, nabízí cenný zdroj pro objevování nových terpenů.

Klíčová slova: terpen syntáza, TPS, mining, Pfam, SUPERFAMILY, doména, terpen

Table of contents

Abbreviations

TPS	terpene synthase
IPP	isopentenyl diphosphate
DMAPP	dimethylallyl diphosphate
GPP	geranyl diphosphate
FPP	farnesyl diphosphate
GFPP	geranylfarnesyl diphosphate
IDS	isoprenyl diphosphate synthase
MTPSL	microbial TPS-like
PTTS	prenyltransferase-terpene synthase
PT	prenyltransferase
HMM	Hidden Markov Model
BLAST	Basic Local Alignment Search Tool
PSI-BLAST	Position-Specific Iterated BLAST
MSA	multiple sequence alignment
SSN	sequence similarity network
ML	maximum likelihood
PLM	protein language model
MDP	maximum diversity problem
TSA	Transcriptome shotgun assembly
t-SNE	t-distributed stochastic neighbor embedding
UMAP	Uniform Manifold Approximation and Projection
PCA	principal component analysis

Introduction

Terpenes and terpenoids represent the largest and most structurally diverse group of natural products (> 80,000 structures known) (Christianson 2017)[1]. These compounds, the biosynthesis of which starts from isoprene units, are present across various domains of life, spanning organisms from plants and fungi to bacteria, marine organisms, and even insects (Tholl et al. 2023). The initial scaffolding step in terpene biosynthesis is carried out by terpene synthases (TPSs), enzymes that catalyze some of the most complex chemical reactions in biology. Remarkably, during the course of a multistep cyclization cascade catalyzed by these enzymes, more than half of the substrate carbon atoms, on average, undergo changes in bonding, hybridization, and stereochemistry (Christianson 2017). Terpene synthases significantly contribute to the high structural diversity of terpenoids by producing a wide array of terpene scaffolds. The pool of terpenoids is then further expanded by activities of other enzymes (such as cytochromes P450s), which modify the initial scaffolds (Zhang et al. 2020; Leferink and Scrutton 2022; Rudolf and Chang 2020).

The importance of terpenoids spans a wide spectrum of fields, including the pharmaceutical industry, food industry, and biofuels (Schwab, Fuchs, and Huang 2013). In the pharmaceutical industry, terpenoids serve as important drugs. For example, artemisinin is a first-line treatment for malaria first discovered by Youyou Tu, who was awarded the Nobel Prize in 2015. Similarly, taxol stands as an important chemotherapeutic drug in the fight against cancer. Additionally, pleuromutilin, first identified in 1951, serves as an antibiotic for livestock and holds promise for human use (Liu, Chen, and Zhang 2023).

The chemical synthesis of terpenoids can be highly complex, and natural extraction often yields low quantities. This underscores the importance of terpene synthases. By genetically engineering host organisms to express these enzymes, it becomes possible to achieve more efficient production (Zerbe and Bohlmann 2015). For instance,

[1] This thesis utilized AI tools, grammarly.com, and OpenAI's ChatGPT for grammar correction and to enhance readability. Their usage did not influence the original research findings or conclusions presented herein.

extracting artemisinin from *Artemisia annua* typically yields 0.1-1% of the plant's dry weight, and chemical synthesis methods suffer from low overall yields and high costs. Contrastingly, utilizing engineered microorganisms such as *Saccharomyces cerevisiae* can boost the overall yields, offering a scalable and cost-effective alternative for commercial production of terpenoids (Zhao et al. 2022; Liu, Chen, and Zhang 2023).

Sequence-guided mining has emerged as a powerful bioinformatics tool to identify terpene synthases across diverse organisms (see Chapter 3). This approach holds promise for discovering novel enzymes and, consequently, novel terpenes with potential applications across various industries, including the pharmaceutical industry.

The goals of this thesis

Through comprehensive analysis of a library of characterized terpene synthases, the goal is to understand the similarities and differences in terpene synthases across different domains/kingdoms of life, as well as different types of TPSs according to their products. Special attention is focused on examining the presence and combinations of conserved domains found in the Pfam (Mistry et al. 2021; Sonnhammer, Eddy, and Durbin 1997) and SUPERFAMILY (Gough et al. 2001) databases.

Using this understanding of terpene synthases, this thesis explores the application of sequence-guided mining techniques in unraveling the landscape of terpene synthases. To the best of my knowledge, this effort represents the largest ever reported TPSs mining using nearly 5.5 billion protein sequences from various sources, resulting in the identification of more than 600 thousand TPS candidates spanning all kingdoms of life.

Furthermore, the thesis also provides a Sequence Similarity Network (SSN), and a large-scale phylogenetic tree for exploration of the TPS candidates, and two scoring mechanisms aimed at prioritizing the candidates for experimental characterization.

The findings of this research enable rapid discovery of novel terpene synthases, which in turn could produce novel terpene compounds with diverse potential applications.

1 Terpene synthases

The basic building blocks of terpenes are activated isoprene units (C5) isopentenyl diphosphate (IPP) and its isomer dimethylallyl diphosphate (DMAPP). Through the action of isoprenyl diphosphate synthases (IDSs), also referred to as prenyltransferases (PTs), DMAPP and variable numbers of IPP are fused to form isoprenyl diphosphates such as geranyl diphosphate (GPP, C10), farnesyl diphosphate (FPP, C15), geranylgeranyl diphosphate (GGPP, C20), and geranylfarnesyl diphosphate (GFPP, C25), which serve as substrates for terpene synthases (Gao, Honzatko, and Peters 2012; Christianson 2017).

Based on the number of isoprene units they contain, terpenes are classified into several types: monoterpenes (C10), sesquiterpenes (C15), diterpenes (C20), sesterterpenes (C25), triterpenes (C30), sesquarterpenes (C35) and tetraterpenes (C40)[2]. Accordingly, TPSs are classified as monoterpene synthases (monoTPSs), sesquiterpene synthases (sesquiTPSs), diterpene synthases (diTPSs), etc. Many terpene synthases are promiscuous and can produce multiple types of terpenes (Gao, Honzatko, and Peters 2012; Christianson 2017). This classification of TPSs is later referred to as TPS types.

Chapter 1.1 reviews the modular protein architecture of TPSs. In Chapter 1.2, the classification of TPSs into two classes, Class I and Class II, is described. These two classes differ by the substrate activation mechanism and also by their structural domain architectures. Finally, Chapter 1.3 compares TPSs from various kingdoms of life, including plants, fungi, bacteria, and some animals, along with a recently discovered TPS from a giant virus (Jung et al. 2023; Tholl et al. 2023).

1.1 Modular architecture

TPSs exhibit a modular architecture of various combinations of three alpha-helical structural domains, α, β, and γ. Ancestral TPS likely consisted of all three domains, with both α domain and βγ domain assembly being catalytically active. Very low sequence and

[2] Longer isoprene polymers (such as natural rubber) exist; however they will not be discussed in this thesis.

structural similarity between α and β domains suggests that α and β evolved independently (Christianson 2017).

The α domain represents the typical Class I terpene synthase fold. This domain forms a bundle comprising 10-12 α-helices, also often referred to as the "isoprenoid" fold (see Figure 1). This is likely due to the same ionization mechanism in Class I TPSs and prenyltransferases (see Chapter 1.2) (Rudolf and Chang 2020; Gao, Honzatko, and Peters 2012). Active α domains contain two highly conserved Mg^{2+}-binding motifs (metal-binding residues are in bold): **D**DXX**D**[3] and NSE/DTE, typically characterized as [**ND**]DXX[**ST**]XXX**E**, although alternative definitions exist (Christianson 2017).

The βγ domain assembly represents the Class II terpene synthase fold, characterized by two α-barrels with an overall dumbbell shape (see Figure 1). The γ domain, positioned between the first and second helices of the β domain, likely originated from a gene duplication event. In some TPSs, such as squalene-hopene cyclases and oxidosqualene cyclases, the γ domain functions as a membrane anchoring component (Christianson 2017). In catalytically active βγ domains, the active site resides at the interface of β and γ and consists of a characteristic Asp-rich DXDD motif unrelated to the DDXXD motif found in the α domain. Unlike Class I motifs, the DXDD motif does not bind Mg^{2+} (Rudolf and Chang 2020).

The modular architecture of TPSs manifests in various combinations, including α, αβ, βγ, αβγ, and αα assemblies with different combinations observed in both Class I and Class II TPSs (Christianson 2017).

[3] This motif is also commonly denoted as **D**DXX**[D/E]**.

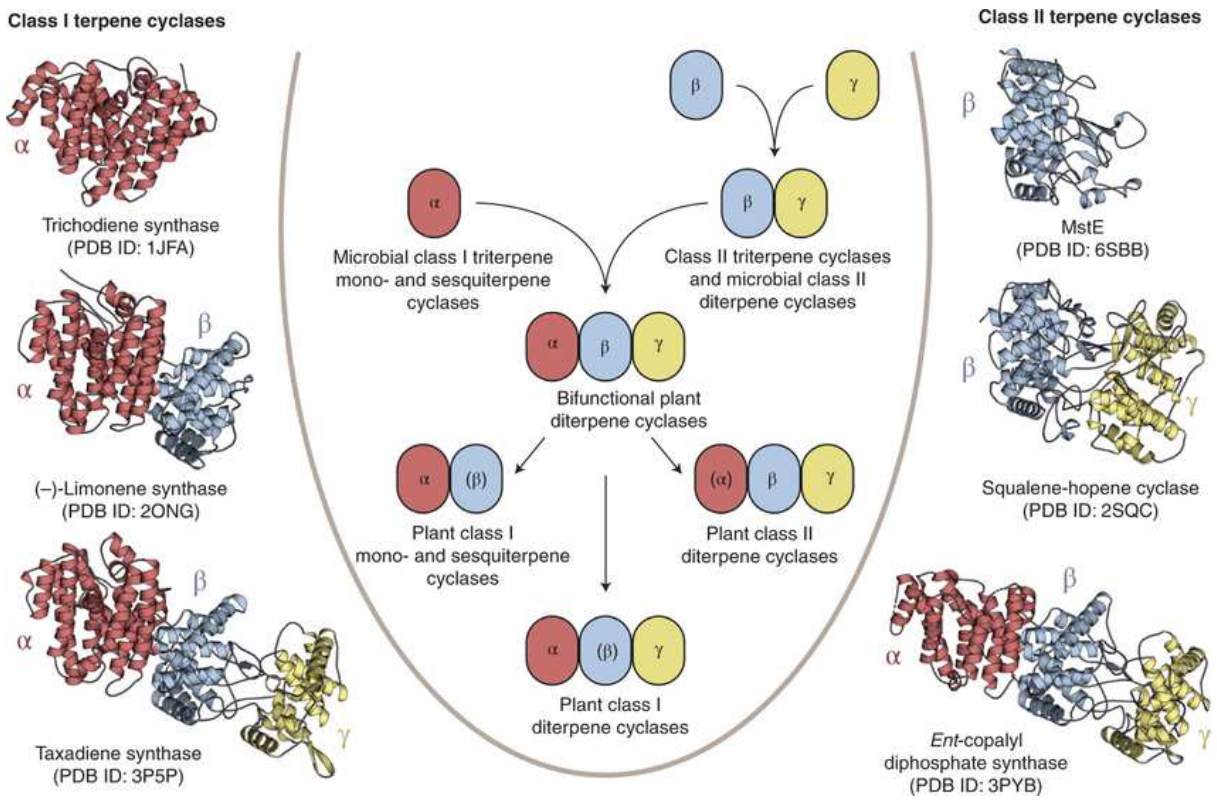

Figure 1. In the center, a schematic representation illustrates the Class I and Class II domain architectures and evolution of TPSs. On the left side, examples of structures of Class I TPSs are provided, with colored domains corresponding to the schematic representations in the center. Similarly, the right side shows examples of Class II TPSs. Adapted from (Moosmann et al. 2020).

1.2 Class I and Class II TPS

With regard to the initiation of catalysis, TPSs can be divided into two classes, which also have different folds, as previously mentioned.

Class I TPSs initiate the reaction by ionization of prenyl-PP substrate. A trinuclear Mg^{2+} cluster, coordinated by the highly conserved DDXXD and NSE/DTE metal-binding motifs situated within the α-domain, binds the prenyl-PP substrate, which provides an electrophilic driving force for the ionization. The same mechanism is used by prenyltransferases. The observed domain architectures include α, αβ, αβγ, or αα, where different architectures are typical for different TPS types and TPSs from different kingdoms. For example, plant TPSs typically have the αβ architecture (where the β

domain is not active), whereas bacterial TPSs typically have only the α domain. Class I TPSs are usually mono-, sesqui-, di-, and sesterTPSs (Christianson 2017).

Class II TPSs employ a distinct catalytic mechanism wherein a conserved aspartic acid in the DXDD motif protonates the terminal carbon–carbon double bond of the prenyl-PP substrate, generating a tertiary carbocation. The active site of a Class II TPSs is located at the interface of β and γ. The observed domain architectures are βγ or αβγ, but recently, β architecture was also observed in cyanobacteria (Moosmann et al. 2020). Plant Class II diTPSs often adopt the αβγ architecture, with the α domain lacking catalytic activity and metal-binding motifs, whereas for bacterial Class II diTPSs, the βγ domain architecture is typical. TriTPSs typically have the βγ fold. Class II TPSs are usually di-, tri-, sester-, and sesquarTPSs (Christianson 2017).

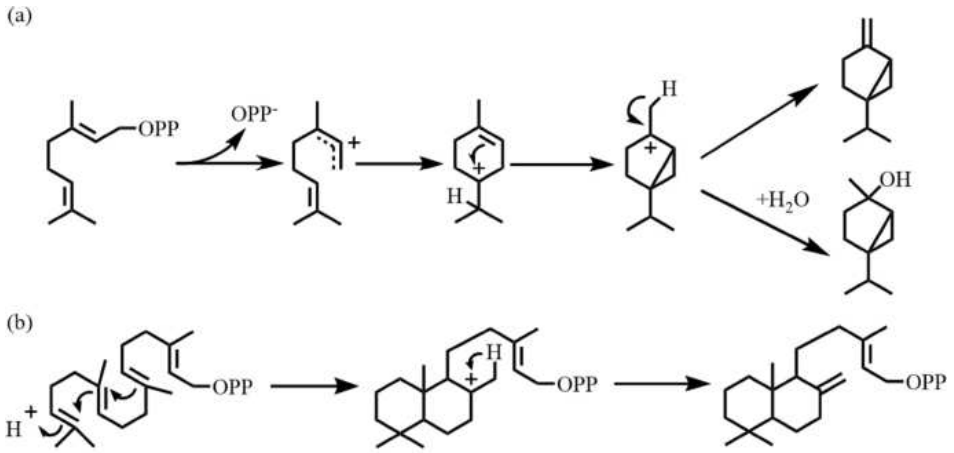

Figure 2. Substrate activation mechanisms in Class I (a) and Class II (b) TPSs. (a) Class I ionization-dependent activation reaction, (b) Class II protonation-dependent activation reaction. Adapted from (Huang et al. 2021).

There are also, although less common, bifunctional TPSs, either Class I-Class I with αα architecture or Class I-Class II with αβγ architecture, same as the ancestral TPS. Plant copalyl diphosphate synthase-kaurene synthase is an example of a Class I-Class II TPS, which probably represents the evolutionary ancestor of all plant TPSs (Christianson 2017).

1.3 TPSs across kingdoms of life

While terpenes from plants are the most extensively studied, terpenes and terpene synthases are found across various kingdoms of life. To date, TPSs have been identified in plants, fungi, bacteria, red algae, amoebae, sponges, corals, and insects. Most recently, TPSs were also surprisingly identified in giant viruses (Jung et al. 2023).

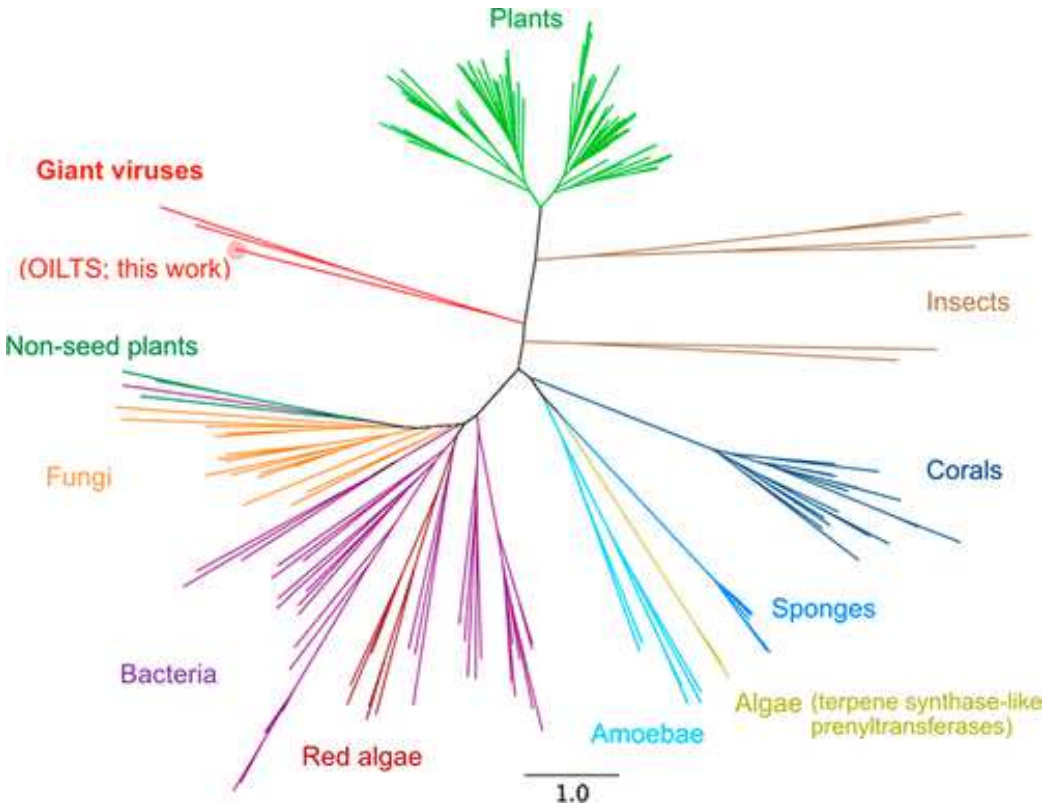

Figure 3. Phylogenetic tree of TPSs from different kingdoms. Adapted from (Jung et al. 2023).

Generally, TPSs can be categorized into plant TPSs, microbial TPSs from Bacteria and Fungi (or microbial-like TPS from other kingdoms), and IDS-like TPSs, with different structural domain architectures observed across various kingdoms (see Figure 4).

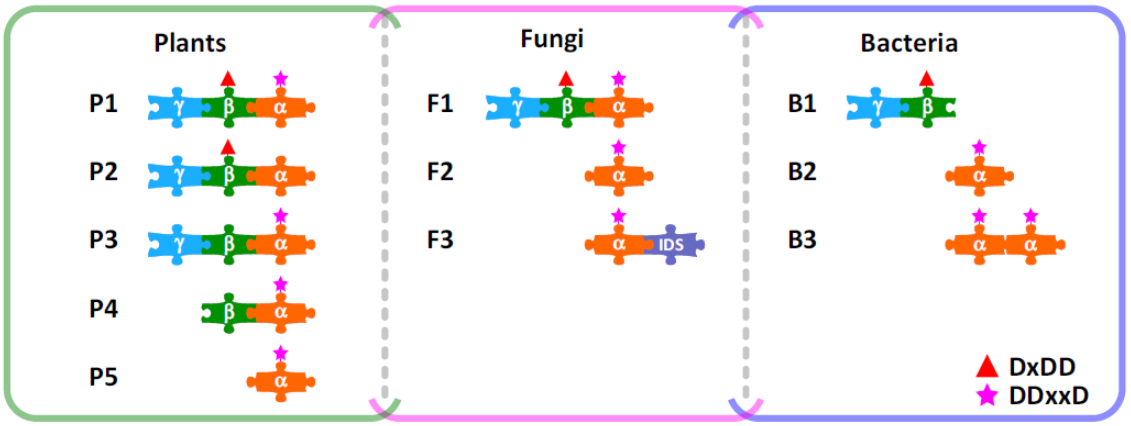

Figure 4. Structural domain architectures and conserved motifs in plants, fungi, and bacteria. Adapted from (Jia et al. 2018).

Plants

In plants, terpenes play essential roles in the production of the phytohormones gibberellins, but the majority of terpenes serve as secondary metabolites with diverse functions. These functions range from defense against herbivores or microbes and attraction of insect pollinators to coping with environmental stress. The remarkable diversity of terpene synthases in plants reflects their complex ecological adaptations (Jia et al. 2018; Boutanaev et al. 2015; S.-Y. Jiang et al. 2019).

Plant genomes exhibit highly variable counts of TPS family genes, usually ranging from 20 to more than 100, with particularly large families observed in dicots. The primary mechanisms driving this expansion are tandem and segmental genome duplications (S.-Y. Jiang et al. 2019; F. Chen et al. 2011).

TPSs in different plant species were primarily identified through sequence similarity and similar gene organization across related taxa (Leferink and Scrutton 2022). The ancestral plant TPS was likely a copalyl diphosphate synthase-kaurene synthase, which is a bifunctional Class I-Class II diTPS. Both Class I and Class II TPSs are common in plants, with typical domain architectures of αβγ or αβ.

Typical plant terpene synthases can be categorized as mono-, sesqui-, di-, or triTPSs. Their lengths typically range from 550 to 800 amino acids, depending on the domain

architecture and subcellular localization. SesquiTPSs are localized in the cytosol, whereas monoTPSs and diTPSs are usually localized in plastids and, therefore, contain additional transit peptides at the N-terminal (Jia et al. 2018; F. Chen et al. 2011).

Beyond seed plants, a distinct group of terpene synthases, known as microbial TPS-like (MTPSLs)/microbial-type TPSs, was identified in non-seed plants, including liverworts, mosses, hornworts, lycophytes, and monilophytes. Phylogenetically and structurally related to bacterial and fungal terpene synthases, MTPSLs likely originated through horizontal gene transfer. These enzymes, mostly monoTPSs or sesquiTPSs, typically contain the α domain architectures and are typically shorter, around 350 amino acids. Non-seed plants also contain typical plant TPSs (Jia et al. 2016, 2018).

Red algae

In red algae, terpene biosynthesis is exclusively mediated by microbial-type TPSs, which, according to phylogenetic analysis, cluster with bacterial TPSs and do not group with microbial-type TPSs from non-seed plants. Microbial-type TPSs in red algae likely originated from an independent horizontal gene transfer event (Wei et al. 2019).

Fungi

In fungi, terpenes play key roles in defense against predators (mycotoxins and phytotoxins), as well as in establishing symbiotic relationships. SesquiTPSs are the most common type of TPSs in fungi, followed by triTPSs and diTPSs (Hage et al. 2023; González-Hernández et al. 2023). These fungal TPSs typically have the α domain architectures or occasionally the αβγ domain architectures and exhibit low sequence similarity to plant TPSs (Jia et al. 2018).

There are Class I TPSs and also Class I-Class II bifunctional TPSs in fungi (Gao, Honzatko, and Peters 2012; Schmidt-Dannert 2015). In addition to typical TPSs, fungi from Dikarya contain a unique class of TPSs called prenyltransferase-terpene synthases (PTTSs) or chimeric TPSs. These enzymes contain a prenyltransferase (PT) domain at the C-terminal and a Class I TPS (TS) domain at the N-terminal. They can directly use IPP and DMAPP to produce di- and sesterterpenes (R. Chen et al. 2021; Jia et al. 2018).

Fungal biosynthetic pathways, including terpenoid biosynthetic pathways, often exhibit gene clustering, simplifying the identification of entire biosynthetic pathways (Quin, Flynn, and Schmidt-Dannert 2014).

Bacteria

Most characterized bacterial TPSs produce one major product, and most characterized bacterial TPSs produce sesquiterpenes (Reddy et al. 2020; Helfrich et al. 2019). Bacterial TPSs exhibit low sequence similarities not only to plant and fungal TPSs but also among themselves. Both Class I and Class II TPSs are known to exist in bacteria (Z. Li et al. 2023). They typically have the α domain architecture, but sometimes also the αα or βγ in diterpene synthases. Recently, a cyanobacterial Class II TPS with only a β domain was reported (Moosmann et al. 2020).

Insects

In insect biology, terpenes typically serve as pheromones for communication, such as mate finding and avoiding predators by signaling danger. (Tholl et al. 2023) Initially, it was believed that terpenes in insects originated from bacterial symbionts, but later, insect TPSs were identified.

Insect TPSs have probably recently evolved from IDSs by their neofunctionalization, and they are often denoted as IDS-like TPSs. These IDS-like TPSs represent a distinct non-canonical group of TPSs (Rebholz et al. 2023).

Other animals

Many animals use terpenes for interactions as they are constituents of pheromones (Tholl et al. 2023). Octocorals (soft corals) are significant contributors to terpenoid diversity in oceans. They mainly produce diterpenes, probably as chemical defenses against predators. Octocoral TPSs form their own clade but are most similar to microbial TPSs. The so far discovered TPSs are typically 400 amino acids long and belong to Class I TPSs (Scesa, Lin, and Schmidt 2022; Burkhardt et al. 2022). Recent discoveries have revealed Class I TPSs also in marine sponges, where it was traditionally thought that terpenes originated from their microbial symbionts. These sponge TPSs

mainly produce sesquiterpenes and contain the DDXXD motif and slightly altered NSE/DTE motif characteristic for Class I TPSs (Wilson et al. 2023). Furthermore, TPSs have been identified in various arthropods, including trombidid mites, millipedes, and arachnids. They likely acquired their TPS genes from microbial sources through horizontal gene transfer (Tholl et al. 2023). TPSs were also identified in social amoebae. These enzymes primarily exhibit sesquiterpene synthase activity and are most closely related to fungal TPSs (X. Chen et al. 2016). Humans and other animals possess lanosterol synthase, a type of oxidosqualene cyclase (triTPS) involved in forming sterols, including intermediates leading to cholesterol (Christianson 2017). Other TPSs have not been identified in animals, and their primary source of terpenoids is diet or symbionts (Tholl et al. 2023).

Giant viruses

Recently, the first TPSs were discovered in giant viruses (giruses), large viruses with extensive genomes. The discovery of TPSs in the giral genomes was surprising, and it remains unclear why these viruses possess TPSs. The giral TPSs form a separate clade but contain the characteristic Class I TPS motifs, DDXXD and NSE/DTE. One giral TPS was experimentally characterized and found to function as both sesquiTPS and monoTPS. It has the α domain architecture and consists of only 278 amino acids. Notably, the genomes of examined giant viruses do not contain prenyltransferases, indicating that the substrates for terpene synthesis would need to be obtained from the host organism (Jung et al. 2023)).

2 Sequence-guided mining

Sequence-guided mining techniques, such as genome mining, aim to extract specific sequences of interest from a target genome, often belonging to particular protein families or superfamilies. In this thesis, my focus lies on the protein superfamily of TPSs. However, rather than targeting a single genome, the approach here involves sequences from various genomic, metagenomic, transcriptomic, and protein databases, a method commonly referred to as global genome mining (Malit, Leung, and Qian 2022).

There are three primary approaches to sequence-guided mining: utilizing sequence alignment algorithms such as BLAST or PSI-BLAST; employing custom profile Hidden Markov Models (HMMs); and leveraging profile HMMs from protein family databases.

The first approach utilizes BLAST (Basic Local Alignment Search Tool), a sequence alignment tool designed to compare query sequences against a sequence database to identify similar sequences (Altschul et al. 1990). This approach proves helpful when the objective is to find homologous protein sequences in closely related species. However, using BLAST is less suited for tasks such as identifying distant homologs with potentially novel functions (Park et al. 1998; Madera and Gough 2002), as in the case of searching for terpene synthases producing novel compounds. There is also an alternative version of BLAST, PSI-BLAST (Position-Specific Iterated BLAST), which aims to find more distant relatives by utilizing a position-specific scoring matrix or profile updated in every iteration of the search (Altschul et al. 1997).

The second approach involves creating and utilizing profile Hidden Markov Models (profile HMMs). Profile HMMs are probabilistic models that can capture conservation patterns within protein families. They are better at handling insertions and deletions, thus enabling the identification of more distant homologs (Krogh et al. 1994; S. R. Eddy 1998). They have proven to be even more sensitive than PSI-BLAST (Madera and Gough 2002). Profile HMMs can be generated and utilized using bioinformatic tools such as HMMER3 (Sean R. Eddy 2023). The profile HMM building process begins with collecting a seed of sequences, followed by multiple sequence alignment (MSA). From the MSA, a

profile HMM is built and then compared with sequences in the databases (Krogh et al. 1994; Sean R. Eddy 2023).

The third approach also utilizes profile HMMs. However, in this case, rather than creating HMMs from scratch using a seed of sequences, protein family databases are leveraged. There are several protein family databases, many collected within the InterPro database (Paysan-Lafosse et al. 2023; Apweiler et al. 2001). The most commonly used database is the Pfam database; another notable database is the SUPERFAMILY database.

Pfam (Mistry et al. 2021; Sonnhammer, Eddy, and Durbin 1997) provides a database of expert-curated protein family alignments and profile HMMs. Pfam is a member of InterPro. Pfam tries to cover as many protein sequences as possible using the fewest possible number of models. The goal is that no models overlap. However, this is not possible in some cases, and it led to the introduction of clans, collections of families with the same evolutionary origin. In the clans, models can overlap (Finn et al. 2006). Each Pfam family is defined by a representative set of sequences (seed), which are aligned to create a profile HMM. An iterative process is then employed to scan a database of sequences *pfamseq* (based on UniProtKB reference proteomes), updating the model with additional sequences aligned to the profile HMM. Finally, each family is annotated with information from literature when available.

SUPERFAMILY 1.75 is another database of expert-curated profile HMMs (Gough et al. 2001). This database is based on SCOP (Structural Classification of Proteins) database 1.75, which classifies protein domains with known 3D structures into a hierarchical system. The first four levels of SCOP classify into classes, folds, superfamilies, and families (Murzin et al. 1995). SUPERFAMILY focuses on the superfamily level, grouping protein domains with structural, functional, and sequence evidence of a common evolutionary ancestor. Each superfamily consists of one or multiple profile HMMs representing the superfamily. SUPERFAMILY 1.75 is a member of InterPro (de Lima Morais et al. 2011). A newer version, SUPERFAMILY 2.0 (Pandurangan et al. 2019), based on a newer version of SCOP, has been developed; however, it is not included in the

commonly used InterProScan suite, and its HMMs are not publicly available for download.

For enzyme mining, there is also a specialized tool known as EnzymeMiner, which is designed to automate the mining process for any enzyme of interest. EnzymeMiner is a web server solution that automates the mining, annotation, and prioritization of candidates for experimental characterization. This method requires input query sequences along with a template file describing essential residues within these sequences. By employing two iteration PSI-BLAST using the query sequences, it mines sequences from the NCBI *nr* database and filters them based on the presence of user-defined essential residues. The next step is the annotation of identified sequences with several tools (Hon et al. 2020). The prioritization step of EnzymeMiner will be described in the subsequent Chapter 2.3.

2.1 HMM domains associated with terpene synthases

2.1.1 Pfam domains

Pfam domains associated with TPSs belong to two Pfam clans, CL0613 (Terp_synthase) and CL0059 (6_Hairpin). Pfam clan CL0613 includes a diverse range of enzyme families involved in terpene biosynthesis, all sharing an alpha-helical core structure (see Figure 1). Pfam clan CL0059 (6_Hairpin) consists of 42 members sharing a common structure composed of 6 helical hairpins.

An overview of the Pfam models associated with TPSs is provided in Table 1, showing only TPS-related models in these clans.

Table 1. IDs and descriptions of Pfam HMM models associated with TPSs.

Pfam clan (ID)	Pfam family	Description
Terp_synthase (CL0613)	**PF03936** (Terpene synthase family, metal binding domain)	Terpene synthase C terminal domain
	PF19086 (Terpene synthase family 2, C-terminal metal binding)	Domain for C-terminal metal binding of class I TPS, especially in bacteria
	PF06330 (Trichodiene synthase)	Family of several fungal trichodiene synthase proteins
	PF00494 (Squalene/phytoene synthase)	Family of squalene synthases and phytoene synthases which share conserved regions
6_Hairpin (CL0059)	**PF01397** (Terpene synthase, N-terminal domain)	Terpene synthase N terminal domain
	PF13243 (Squalene-hopene cyclase C-terminal domain)	Squalene-hopene cyclase catalyses the cyclisation of squalene into hopene.This family is the C-terminal domain
	PF13249 (Squalene-hopene cyclase N-terminal domain)	Squalene-hopene cyclase catalyses the cyclisation of squalene into hopene.This family is the N-terminal domain

Additionally, in Figure 5, the taxonomic distribution of sequences captured by Pfam models in the InterPro database is presented. For instance, domains PF03936 and PF01397 predominantly capture sequences from plants, while PF06330 primarily captures fungal sequences. PF19086 mainly captures microbial sequences. PF00494, PF13243, and PF13249 capture sequences from both Eukarya and Bacteria.

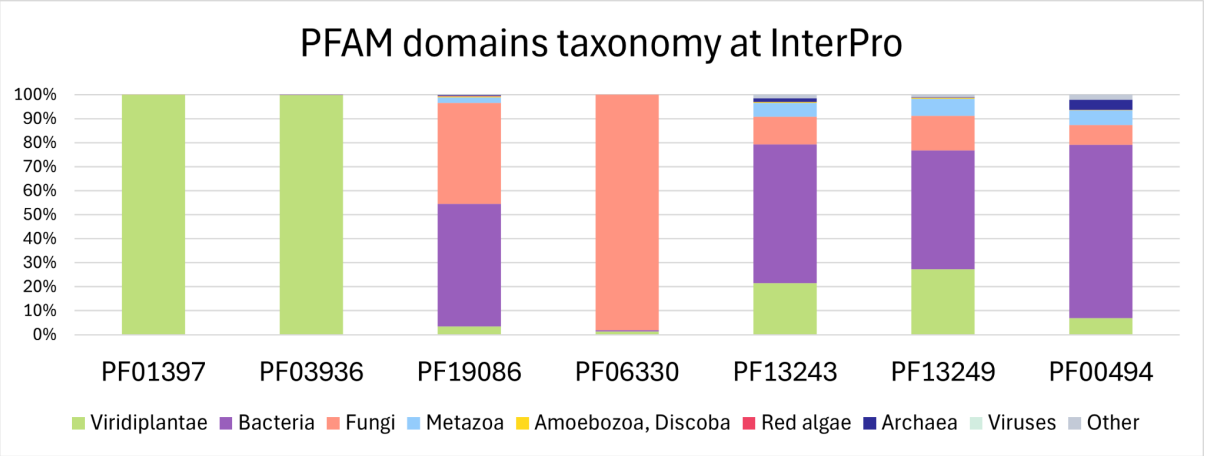

Figure 5. Taxonomic distribution of sequences in InterPro captured by Pfam HMM models.

2.1.2 SUPERFAMILY domains

TPSs are associated with two superfamilies in the SUPERFAMILY database: the Terpenoid synthases superfamily (48576) and the Terpenoid cyclases/Protein prenyltransferases superfamily (48239). The Terpenoid synthases superfamily consists of 5 families, and the Terpenoid cyclases/Protein prenyltransferases superfamily consists of 4 families. Moreover, each family then consists of 1 or more HMM models. Table 2 summarizes the superfamilies. It's worth noting that while some models represent terpene synthases, others encompass additional enzymes involved in terpene biosynthesis.

Table 2. IDs and descriptions of SUPERFAMILY superfamilies, families, and HMM models associated with TPSs.

Superfamily name (SCOP ID)	Family name (SCOP ID)	HMM ID (ASTRAL seed ID)	Seed organisms (ASTRAL seed ID)	Seed description
Terpenoid synthases superfamily (48576)	Isoprenyl diphosphate synthases (48577)	**43373**	Staphylococcus aureus (d1rtra_)	Farnesyl diphosphate synthase (geranyltranstransferase) domain
		49855	Chicken (Gallus gallus) (d1ubya_)	Farnesyl diphosphate synthase (geranyltranstransferase) domain
		43350	Escherichia coli (d1rqja_)	Farnesyl diphosphate synthase (geranyltranstransferase) domain
		44612	Thermotoga maritima (d1v4ea_)	Octoprenyl-diphosphate synthase domain
		54583	Human (Homo sapiens) (d2q80a1)	Geranylgeranyl pyrophosphate synthetase domain
	Squalene synthase (48580)	**46658**	Human (Homo sapiens) (d1ezfa_)	Squalene synthase domain
	Terpenoid cyclase C-terminal domain (48583)	**53355**	Tobacco (Nicotiana tabacum) (d5eaua2)	5-Epi-aristolochene synthase domain
		48261	Garden sage (Salvia officinalis) (d1n1ba2)	(+)-bornyl diphosphate synthase domain
	Aristolochene/pentalene synthase (48586)	**46340**	Fungus (Penicillium roqueforti) (d1di1a_)	Aristolochene synthase domain
		48806	Streptomyces sp., UC5319 (d1ps1a_)	Pentalenene synthase domain
	Trichodiene synthase (69113)	**47573**	Fusarium sporotrichioides (d1jfaa_)	Trichodiene synthase domain

Terpenoid cyclases/Protein prenyltransferases superfamily (48239)	Terpenoid cyclase N-terminal domain (48240)	**53354**	Tobacco (Nicotiana tabacum) (d5eaua1)	5-Epi-aristolochene synthase domain
		41184	Garden sage (Salvia officinalis) (d1n1ba1)	(+)-bornyl diphosphate synthase
	Terpene synthases (48243)	**53306**	Alicyclobacillus acidocaldarius (d2sqca2)	Squalene-hopene cyclase domain
		50379	Human (Homo sapiens) (d1w6ka1)	Lanosterol synthase domain
		53305	Alicyclobacillus acidocaldarius (d2sqca1)	Squalene-hopene cyclase domain
		50380	Human (Homo sapiens) (d1w6ka2)	Lanosterol synthase, middle domain
	Protein prenyltransferases (48246)	**46282**	Rat (Rattus norvegicus) (d1d8db_)	Protein farnesyltransferase, beta-subunit domain
		48283	Rat (Rattus norvegicus) (d1n4qb_)	Protein farnesyltransferase, beta-subunit domain
	Complement components (48251)	**35832**	Human (Homo sapiens) (d1c3da_)	Thio-ester containing domain (TED) from Complement C3, aka C3d or C3dg
		49012	Rat (Rattus norvegicus) (d1qqfa_)	Thio-ester containing domain (TED) from Complement C3, aka C3d or C3dg
		47273	Human (Homo sapiens) (d1hzfa_)	C4adg fragment of complement factor C4a domain

In InterPro, only the superfamily level is considered, meaning that a superfamily is assigned to a sequence if it is identified by any HMM model within that superfamily. Figure 6 illustrates the taxonomic distribution of sequences captured by superfamilies in the InterPro database. Compared with Pfam domains, there appears to be a larger proportion of bacterial and metazoan sequences. However, it is important to note that this data encompasses the entire superfamily and may not exclusively represent TPSs.

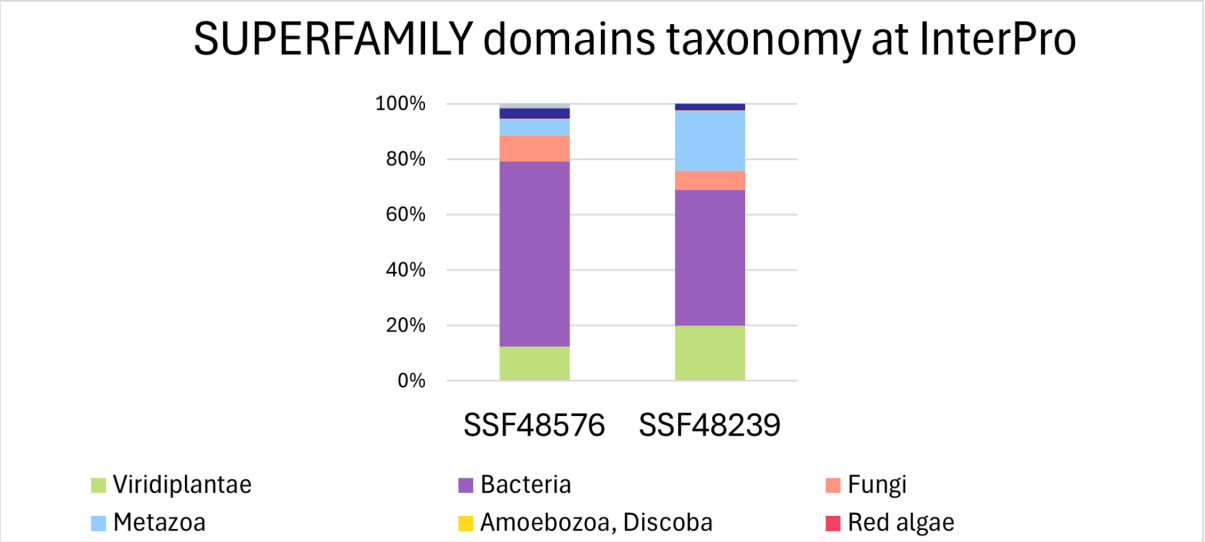

Figure 6. Taxonomic distribution of sequences in InterPro captured by the SUPERFAMILY HMM models. This data encompasses the entire superfamily and may not exclusively represent TPSs.

2.2 Analyzing the mined sequence space of protein families

When analyzing mined sequences, annotating them with relevant metadata is essential to gather as much information as possible to effectively explore the sequence space of the protein family or superfamily. This typically includes obtaining information such as the source organism (using its NCBI Taxon ID) and its kingdom, the presence of specific functional motifs and HMM domains, and annotations of the most similar annotated characterized sequences. Additionally, the sequences' structural similarity can be explored by obtaining or predicting 3D structures.

To visualize and analyze the sequence space, a combination of characterized and annotated sequences is used along with the metadata-annotated set of mined sequences. Commonly used techniques include phylogenetic trees and sequence similarity networks (SSNs). Additionally, projections of protein embeddings of the sequences are sometimes used.

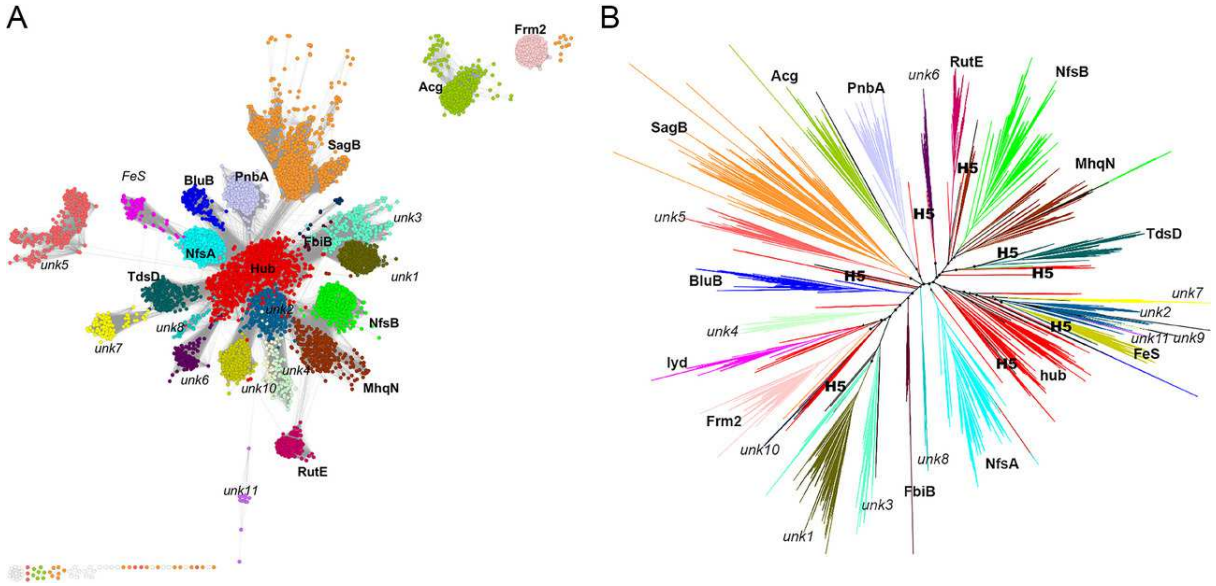

Figure 7. Comparison of SSN (A) and phylogenetic tree (B) on the same data. Adapted from (Copp et al. 2019; Akiva et al. 2017).

2.2.1 Phylogenetic trees

Phylogenetic trees serve as a traditional tool for visualizing evolutionary relationships among sequences within a protein family or superfamily. The output from phylogenetic analysis yields a set of nested relationships between the input sequences (aligned in an MSA), which are then visualized as a phylogenetic tree. In the phylogenetic tree, sequences are represented as leaves, and their connections represent the relationships (Copp et al. 2019). Phylogenetic trees can be used to infer functions of proteins based on characterized orthologs (Brown and Sjölander 2006). Exploring distant clades without characterized sequences can increase the knowledge about the given protein family,

The process involves building an MSA, which is subsequently used for phylogenetic tree construction. From the sequence conservation, MSAs can detect features that determine function, such as key catalytic residues (Copp et al. 2019). Various algorithms, such as Maximum Likelihood (ML) or Bayesian methods, can be used to construct phylogenetic trees.

Constructing phylogenetic trees can be challenging, especially with large or/and diverse sets of sequences. Building an MSA with such a set of sequences can be prone to errors (Atkinson et al. 2009). Furthermore, inference of function relies on the assumption that the phylogenetic tree topology is correct, which may not always be true and can lead to errors. Events such as gene duplication and neofunctionalization, and varying evolutionary rates can even further complicate the construction of a correct phylogenetic tree as it may violate the assumptions of evolutionary models (Brown and Sjölander 2006).

Despite these challenges, phylogenetics remains a key method for the exploration of protein superfamilies. Compared to other methods, its main strength is better capturing of key functional features. An example of a helpful tool for visualizing large phylogenetic trees with several annotations is iTOL (Letunic and Bork 2021).

2.2.2 Sequence similarity networks (SSN)

Sequence similarity networks (SSN) have emerged as a popular alternative to phylogenetic trees for visualizing sequence relationships within protein families. In SSNs, sequences are represented as nodes, and connections between nodes indicate sequence similarity above a defined threshold (Atkinson et al. 2009). Unlike phylogenetic trees, SSNs do not rely on evolutionary models, and therefore, they cannot be used to infer evolutionary relationships (Copp et al. 2019).

SSNs are relatively easy to read, even with large datasets, and provide an effective view of sequence similarity relationships. After overlaying with annotations, SSNs can reveal clusters of sequences with similar functions (isofunctional clusters), as well as unexplored clusters that may contain sequences with novel functions (Atkinson et al. 2009).

However, SSNs also pose challenges, particularly regarding the selection of similarity thresholds since there is no universal threshold. The distance metric, based on sequence similarity, can be misleading. In some cases, members of a protein family with the same function can have sequence similarity of less than 20%, while in some other cases, sequences with over 90% similarity can have distinct functions (often due to a single residue mutation). Therefore, the separation of sequences into clusters based on sequence similarity does not guarantee isofunctionality(Copp et al. 2018).

SSNs can be created manually by running all-vs-all BLAST, as described in detail in (Copp et al. 2018), or using EFI-EST, a web-based platform for generating SSNs (Oberg, Zallot, and Gerlt 2023; Rémi Zallot, Oberg, and Gerlt 2019). SSNs are typically visualized in Cytoscape (Shannon et al. 2003).

2.2.3 Protein sequence embeddings

The utilization of protein sequence embeddings from protein language models (PLMs) is gaining popularity for various tasks in computational biology. Remarkably, PLMs have demonstrated the ability to learn structural, evolutionary, and biochemical properties solely from sequences (Rives et al. 2021; Elnaggar et al. 2022).

In protein embeddings, each sequence is represented as a single numerical vector (an *embedding*), thereby occupying a single point in a high-dimensional space. Sequences with similar representations are mapped to nearby points. It was shown that remote homologs with similar structures but divergent sequences cluster in the representation space (Rives et al. 2021).

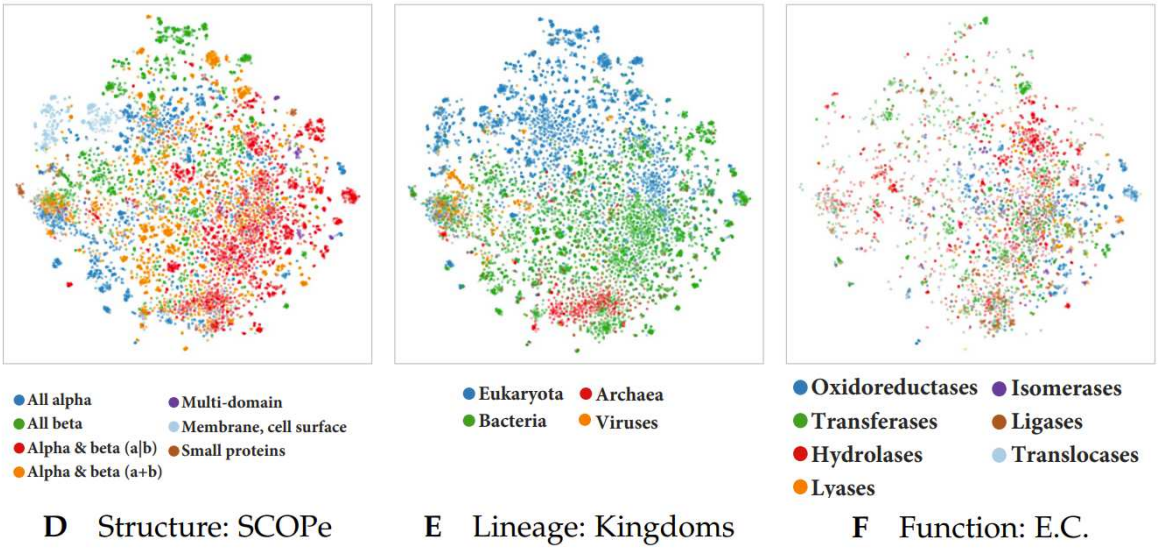

Figure 8. t-SNE projection of protein embeddings annotated with SCOPe classes (D), kingdoms (E), and function (F). Adapted from (Elnaggar et al. 2022).

The Transformer architecture (Vaswani et al. 2017), the most successful machine learning model architecture of PLMs, incorporates an attention mechanism that can be analyzed and visualized. For instance, in the case of the zinc-finger protein, it was shown that attention heads learned to detect the zinc-finger motif, a motif crucial for DNA and RNA binding. This motif consists of residues distant in the sequence but close in structure (Elnaggar et al. 2022).

To explore the sequence space, embeddings can be projected to lower dimensions (2D or 3D) using dimensionality reduction algorithms such as t-distributed stochastic neighbor embedding (t-SNE) (Maaten and Hinton 2008), Uniform Manifold Approximation and Projection (UMAP) (McInnes, Healy, and Melville 2018), or principal component analysis (PCA).

2.3 Candidate selection for novelty discovery

After sequence mining, the goal can often be to select a diverse collection of enzymes for experimental characterization. Effective exploration of the sequence space is a difficult task. There are no standardized methods, and the task usually requires manual expert decisions.

The problem of effective exploration of the sequence space can be viewed as an MDP (maximum diversity problem), which is an NP-hard problem. Heuristic algorithms such as tabu search or k-medoids address the MDP. Still, the development of algorithms for the automated selection of diverse sets of proteins remains mostly unexplored due to the difficulty of this problem. One example of the tabu search algorithm adjusted to the selection of diverse enzymes from an enzyme family is described in (Atallah et al. 2024).

A more established approach, as mentioned earlier, involves manual expert decision-making. This decision-making process can be guided by annotated phylogenetic trees, sequence similarity networks, and other relevant properties depending on the specific problem (such as prioritizing a particular family of species or certain biochemical properties).

In the context of phylogenetic trees, particular interest is directed towards clades lacking characterized sequences, offering potential opportunities for novelty discovery. Successful examples of investigations of such clades include those conducted by (Yamada et al. 2015), which led to the discovery of novel terpenes, (S. Chen, Zhang, and Zhang 2022), which led to the discovery of novel polyketides, and (Kang and Brady 2014), which led to the discovery of novel pentangular polyphenols. Nonetheless, it's crucial to acknowledge that not all divergent enzymes catalyze new chemical reactions (Malit, Leung, and Qian 2022). Some enzymes may produce known products despite their evolutionary distance from characterized sequences, while other distant enzymes may be pseudogenes or possess entirely different functions.

Similarly, in SSNs, particular interest is directed towards clusters without characterized sequences from which the candidates can be chosen. Several explorations across various protein families have been summarized by (Remi Zallot, Oberg, and Gerlt 2021). Notably,

an SSN analysis of bacterial TPSs also led to the discovery of novel terpenes (Hu et al. 2023). Although SSNs have proven successful in numerous cases, it is important to be aware of the fact that the sequence similarity distance metric is very simple and does not guarantee that unexplored clusters will have new functions.

As previously introduced, EnzymeMiner aims to automate the process of mining and selecting novel enzymes. However, the step of selecting candidates from a results table remains manual. Results are presented across multiple tables to guide the selection by, for example, covering sequences from diverse phyla focusing on extremophilic organisms, covering sequences from different SSN clusters, covering sequences with or without additional Pfam domains, covering sequences annotated in NCBI with disease annotations or sequences with known 3D structures. In addition, results can also be filtered to exclude proteins with predicted low solubility, proteins with transmembrane regions, or functionally annotated proteins present in the Swiss-Prot knowledgebase (Hon et al. 2020).

3 Literature overview on TPS sequence mining

There is a large body of previous research on sequence-guided mining of terpene synthases from a particular genome of interest, especially in plants. In addition, there are also various publications where TPS sequences were mined from a larger collection of genomes, often focusing on a specific kingdom. This chapter attempts to summarize the methods used in these publications.

In summary, the methods described in Chapter 2 were used by different research teams, ranging from BLAST and custom HMMs to Pfam models. Furthermore, in most publications, mined sequences were also inspected for the presence of functional motifs along with the expected sequence lengths. Most research teams used phylogenetic trees to explore the sequence space, but there are also examples of using SSNs in Fungi and Bacteria.

Plants

The exploration of terpene synthases in plants has been extensive, often focusing on specific species or families of species. A common approach involves scanning genomes or transcriptomes using Pfam domains PF01397 and PF03936, or using BLAST with query TPS sequences from related species. Examples of such studies include investigations in *Arabidopsis thaliana* (Aubourg, Lecharny, and Bohlmann 2002), maize (Sun et al. 2023), mint (Z. Chen et al. 2021), *Vitis vinifera* (Martin et al. 2010), *Eucalyptus globulus* (Külheim et al. 2015), *Pinaceae* (K. Jiang et al. 2023), or *Citrus sinensis* (Alquézar et al. 2017). In many studies, the researchers have also focused on phylogenetic analyses of plant TPSs (F. Chen et al. 2011; Jia et al. 2022; Yan et al. 2023).

In a few publications, researchers mined diverse collections of plant genomes. (Jia et al. 2016) focused on microbial-like TPSs in plants, mining over 1000 species from the 1KP database using Pfam models PF03936 and PF01397. They identified 712 microbial-like genes in non-seed plants and conducted a phylogenetic analysis in which the plant microbial-like TPS genes clustered with bacterial or fungal TPSs. (Boutanaev et al. 2015) mined 17 plant genomes using genome annotations and BLAST, using monocot and dicot sequences from UniProt annotated as terpene synthases as the query. (S.-Y. Jiang et al.

2019) utilized Pfam domains PF01397 and PF03936 to mine genomes of 50 lower and higher plant species from the Phytozome database, filtering for “full-length” sequences containing both domains, followed by phylogenetic analysis. They further investigated the similarity between IDS proteins and TPSs in plants. For this, they used the Pfam domain PF00348 to mine IDSs, of which 4.36% could also be found using PF03936 with low e-value, indicating some degree of similarity between them. (Yan et al. 2023) mined 74 plant species from diverse plant lineages also utilizing Pfam domains PF01397 and PF03936 to study the evolution of TPSs in plants. In the mining, they identified 3,600 “full-length” sequences containing both domains, 513 sequences containing only the PF01397 domain, and 1,049 sequences containing only the PF03936 domain. Comparing the mined plant TPSs with mined TPSs from microbes, they hypothesize that fusion of PF03936 and PF01397 occurred in an ancestral land plant, which likely acquired both domains independently from microbes through horizontal gene transfer.

Red algae

(Wei et al. 2019) mined seven genomes and 34 transcriptomes of red algae using Pfam models PF01397, PF03936, and PF06330, resulting in the discovery of three microbial-like TPSs in red algae.

Fungi

In fungi, several publications have explored TPSs across multiple genomes. Some studies, such as those by (Quin, Flynn, and Schmidt-Dannert 2014; Zhang et al. 2020), used BLAST for mining. (R. Chen et al. 2021) focused on chimeric PTTSs using Pfam domains PF03936 and PF00348 (Polyprenyl synthetase) for mining. (Hage et al. 2023) used BLAST to mine bacterial sesquiTPS, from which they created four HMMs, which were further updated with new hits. They also concluded that the Pfam model PF19086 better captures fungal TPSs in comparison to PF03936.

To analyze the sequence space, (Quin, Flynn, and Schmidt-Dannert 2014; Hage et al. 2023; R. Chen et al. 2021) used phylogenetic trees. (Zhang et al. 2020) used SSN and found out that the characterized fungal TPSs from Ascomycota are actually located in minor clusters, while the major clusters in the SSN did not contain characterized species and, therefore, present opportunities for the discovery of novel TPSs.

Bacteria

The exploration of bacterial TPSs is gaining increased attention, with multiple recent studies delving into the rich diversity of TPSs present within bacterial genomes.

(Cane and Ikeda 2012; Yamada et al. 2015) used the Pfam model PF03936 to mine bacterial genomes and to iteratively construct their own HMMs from the identified sequences. Similarly, the Pfam model PF03936 was also used by (Reddy et al. 2020). (Hu et al. 2023) used the Pfam model PF19086, which was added to Pfam recently and targets the bacterial TPSs better than PF03936. ((Z. Li et al. 2023) utilized the InterPro database to obtain all bacterial sequences categorized as "terpene cyclase-like 2" (IPR034686). Lastly, (Chhalodia et al. 2023) used BLAST to mine TPSs from bacterial genomes.

Sequence space exploration involved both phylogenetic trees (Yamada et al. 2015; Reddy et al. 2020; Chhalodia et al. 2023) and SSNs (Z. Li et al. 2023; Hu et al. 2023), focused on selecting candidates from clades/clusters separated from the clades/clusters containing characterized TPSs.

Animals

(X. Chen et al. 2016) employed Pfam models PF03936, PF01397, and PF06330 to scan 168 well-annotated genomes of non-plant/non-fungus eukaryotes, discovering TPSs in Amoebozoa. Recent studies focused on octocorals and have employed the Pfam model PF19086 (Scesa, Lin, and Schmidt 2022) and a custom HMM from bacterial and fungal TPSs (Burkhardt et al. 2022) to successfully mine octocoral TPSs. (Wilson et al. 2023) also built a custom HMM using a diverse set of characterized fungal, bacterial, plant, and coral TPSs to mine TPSs from sponges.

All of these studies analyzed their mined TPS sequences in the context of TPSs from other kingdoms using phylogenetic trees, which highlighted that all discovered TPSs form separate clades.

Giant viruses

(Jung et al. 2023) utilized the Pfam model PF19086 to uncover TPSs within giant viruses, discovering 3 TPSs. These TPSs were contextualized alongside TPSs from different kingdoms in a phylogenetic tree, forming their own distinct clade, as depicted in Figure 3.

4 Data and methods

4.1 Data

In this section, a description of the utilized data will be presented, including a manually curated database of characterized terpene synthases (TPS db), large-scale repositories of sequences used for the mining, and protein family profile HMM databases.

4.1.1 TPS database - a curated database of characterized terpene synthases

In the Pluskal lab at IOCB Prague, a database of characterized terpene synthases has been manually collected and curated, hereafter referred to as the TPS database or TPS db (Samusevich et al. 2024). There is an ongoing effort in the group to manually collect data from all published experimentally characterized terpene synthases, and several lab members, including the author, have contributed to the curation process. This database represents the "ground truth" regarding our knowledge of terpene synthases.

The dataset consists of entries representing terpene synthase reactions, each containing several attributes. For this thesis, only selected attributes are detailed, including:

- the ID of the terpene synthase (typically a UniProt ID, but alternatively an NCBI ID or other identifier if the sequence is absent in the UniProt database),
- terpene synthase name,
- protein sequence,
- species,
- kingdom,
- type (mono/di/sesqui,...),
- product name,
- ChEBI ID of the product,
- fragment status (boolean),
- experimentally characterized status (boolean), and

- publication details.

This dataset was and continues to be constructed by gathering information on published terpene synthases with experimentally characterized reaction products. Entries were initially obtained through a manual review of UniProt entries assigned under protein family categories in Table 3. This manual approach was necessary as some entries within the families were assigned to the families based on similarity but lacked experimental characterization. Furthermore, entries were supplemented from additional resources, such as recent publications where data was not yet present in UniProt.

Table 3. TPS protein family categories in UniProt.

Protein Family Name
Lycopene beta-cyclase family
Phytoene/squalene synthase family
Phytoene/squalene synthase family, CrtM subfamily
Terpene cyclase/mutase family
Terpene synthase family
Terpene synthase family, 2-methylisoborneol synthase subfamily
Terpene synthase family, Tpsa subfamily
Terpene synthase family, Tpsb subfamily
Terpene synthase family, Tpsc subfamily
Terpene synthase family, Tpsd subfamily
Terpene synthase family, Tpse subfamily
Terpene synthase family, Tpsf subfamily
Terpene synthase family, Tpsg subfamily
Trichodiene synthase family

When writing this thesis, the dataset comprised a total of 2515 reaction entries corresponding to 1323 unique proteins. However, it is important to note that the dataset originally also includes IDSs, sequences not yet experimentally characterized (usually IDSs), fragmented sequences, and some incomplete entries. Sequences from these categories were filtered out, and only the remaining sequences were utilized, forming a dataset of 1198 entries corresponding to 1125 proteins. Both the original TPS db and the filtered TPS db datasets are available as attachments (See A.2). Overview of the filtered TPS db dataset is described in Chapter 4.2.

4.1.2 Protein sequence databases

This section provides a description of six large-scale sequence repositories (databases) used for the mining. Each database is described below, along with a summary table of database sizes and corresponding download links. The complete collection consists of nearly 5.5 billion sequences, although there is some redundancy in the sources. To the author's best knowledge, this is the largest terpene synthase mining effort conducted up to this day.

1KP (Carpenter et al. 2019; One Thousand Plant Transcriptomes Initiative 2019) is an initiative that collects data from over 1,000 plant transcriptomes (1KP). Selected species encompass a broad diversity of the plant taxonomy (One Thousand Plant Transcriptomes Initiative 2019).

TSA (Transcriptome Shotgun Assembly Sequence Database) (Sayers et al. 2023), a member of the GenBank database, contains assembled transcriptomic data. The nucleotide sequence data was obtained from the NCBI FTP GenBank site. However, since the transcriptomic sequences are nucleotide sequences, the resulting protein sequences were predicted using the TransDecoder tool (Haas 2022).

UniParc (UniProt Archive) (UniProt Consortium 2023) is a non-redundant database collecting sequences from various sources, including UniProt, GenBank, Ensembl, EnsemblGenomes, PDB, RefSeq, and more. It removes redundant entries by assigning each unique protein sequence a UPI identifier.

Phytozome (Goodstein et al. 2012) is a database of green plant genomes and associated data, including amino acid FASTA files of all gene coding sequences. The latest release, version 13, contains 395 assembled and annotated genomes.

MGnify (Richardson et al. 2023) is a metagenomics database containing data from 297 various environments, including, for example, marine, soil, microbiome, and host-associated samples.

BFD (Big Fantastic Database) (Jumper et al. 2021) is a large database of protein sequences from UniProt, MetaClust, and Soil Reference Catalog and the Marine

Eukaryotic Reference Catalog, which were clustered using Linclust/MMseqs2. This database was originally created for developing AlphaFold (Jumper et al. 2021).

Table 4. Summary of protein sequence databases utilized for TPS mining. * indicates the number of protein sequences predicted with TransDecoder.

Database	Size (GB)	Number of protein sequences	Plant only	Download source
1KP	8	25 241 940	Yes	ftp://parrot.genomics.cn/gigadb/pub/10.5524/100001_101000/100627/assemblies/
TSA	130	194 875 849*	No	ftp://ftp.ncbi.nlm.nih.gov/genbank/tsa/ (folders G, H, I)
Phytozome	7	11 952 181	Yes	https://genome.jgi.doe.gov/portal/ (v9-v12 using API) https://data.jgi.doe.gov/refine-download/phytozome (v13 manually)
UniParc	226	543 244 145	No	https://ftp.expasy.org/databases/uniprot/current_release/uniparc/uniparc_active.fasta.gz
MGnify	597	2 477 479 951	No	https://ftp.ebi.ac.uk/pub/databases/metagenomics/peptide_database/current_release/
BFD	1556	2 204 390 010	No	https://bfd.mmseqs.com/bfd_metaclust_clu_complete_id30_c90_final_seq.sorted_opt.tar.gz

4.1.3 Protein domain databases

For analysis of domain architectures and mining, the Pfam and SUPERFAMILY databases of domain profile HMMs were used. All HMMs were considered, and each model's precision was estimated using the SwissProt database protein family annotation. The model was used to mine the SwissProt database, and for each mined sequence, the protein family annotation was retrieved. Depending on whether the protein family belonged to one of the “allowed” terpene synthase families (families in Table 3) or not, it was assigned as either true positive (TP) or false positive (FP) respectively. Then, the precision estimate was calculated as: $precision = TP/(TP + FP)$.

If the estimated precision was high enough, the model was selected for the subsequent large-scale mining effort. The collections of the selected domains are denoted as TPS Pfam db and TPS SUPERFAMILY db.

4.1.3.1 Pfam

All possible TPS Pfam models described in Chapter 2.1.1 were included in TPS Pfam db and used for architecture analysis and mining. Table 5 shows the precision estimate for each of them. Pfam models PF01397, PF03936, and PF19086 have all estimated precision over 90%, and except Pfam model PF13249, with an estimated precision of 69%, all Pfam models have an estimated precision higher than 70%.

Table 5. Precision estimates for Pfam HMM models

Domain ID	Description	Precision estimate
PF00494	Squalene/phytoene synthase	0.77
PF01397	Terpene synthase, N-terminal domain	0.95
PF03936	Terpene synthase family, metal binding domain	0.93
PF06330	Trichodiene synthase	0.84
PF13243	Squalene-hopene cyclase C-terminal domain	0.76
PF13249	Squalene-hopene cyclase N-terminal domain	0.69
PF19086	Terpene synthase family 2, C-terminal metal binding	0.91

4.1.3.2 SUPERFAMILY

For SUPERFAMILY HMMs, precision estimates fluctuated more among the models, and only those models with precision estimates higher than 70% were included in TPS SUPERFAMILY db. In the table below, selected models are highlighted in bold. The low precision estimates of some models are likely caused by the fact that the models represent other enzyme families from the terpene biosynthetic pathway, such as prenyltransferases, as indicated by the model descriptions. However, in the case of certain models, particularly those within Superfamily 48239, the descriptions themselves do not raise suspicions that these models should not represent terpene synthases. Nevertheless, these models demonstrated low precision estimates, and their inclusion could potentially result in a significant number of false positives in the large-scale mining.

Table 6. Precision estimates for SUPERFAMILY HMM models. Selected domains with precision estimates of over 70% are highlighted in bold.

Superfamily (SCOP ID)	Family	Domain ID	Description	Precision estimate
Terpenoid synthases superfamily (48576)	Isoprenyl diphosphate synthases	43373	Farnesyl diphosphate synthase (geranyltranstransferase) domain	0.13
		49855	Farnesyl diphosphate synthase (geranyltranstransferase) domain	0.19
		43350	Farnesyl diphosphate synthase (geranyltranstransferase) domain	0.11
		44612	Octoprenyl-diphosphate synthase domain	0.21
		54583	Geranylgeranyl pyrophosphate synthetase domain	0.22
	Squalene synthase	46658	Squalene synthase domain	0.58
	Terpenoid cyclase C-terminal domain	**53355**	**5-Epi-aristolochene synthase domain**	0.84
		48261	**(+)-bornyl diphosphate synthase domain**	0.85
	Aristolochene/pentalene ne synthase	**46340**	**Aristolochene synthase domain**	0.71
		48806	**Pentalenene synthase domain**	0.72
	Trichodiene synthase	**47573**	**Trichodiene synthase domain**	0.77
Terpenoid cyclases/Protein prenyltransferases superfamily (48239)	Terpenoid cyclase N-terminal domain	**53354**	**5-Epi-aristolochene synthase domain**	0.90
		41184	**(+)-bornyl diphosphate synthase**	0.90
	Terpene synthases	53306	Squalene-hopene cyclase domain	0.51
		50379	Lanosterol synthase domain	0.66
		53305	Squalene-hopene cyclase domain	0.56
		50380	Lanosterol synthase, middle domain	0.61
	Protein prenyltransferases	46282	Protein farnesyltransferase, beta-subunit domain	0.59

		48283	Protein farnesyltransferase, beta-subunit domain	0.59
	Complement components	35832	Thio-ester containing domain (TED) from Complement C3, aka C3d or C3dg	0.64
		49012	Thio-ester containing domain (TED) from Complement C3, aka C3d or C3dg	0.58
		47273	C4adg fragment of complement factor C4a domain	0.56

4.2 Analysis of the characterized terpene synthases

The initial step, as previously mentioned in Chapter 4.1.1, involved filtering of the TPS db, wherein entries identified as fragments or IDSs, entries lacking experimental characterization, and entries with missing information were filtered out. This filtration process resulted in a curated dataset comprising 1125 unique TPSs. Figure 9A illustrates the taxonomic distribution within this dataset, revealing that plant TPSs make up the largest part, accounting for 70% of all TPSs. Following behind are fungal (17%) and bacterial (6%) TPSs. The remaining part of the dataset consists of TPSs from animals (including corals, insects, and marine sponges), amoebozoa, and red algae. Figure 9B shows that among the characterized TPSs, sesquiTPSs appear as the most abundant, followed by mono-, di-, and triTPSs. SesterTPSs and tetraTPSs are rare. Moreover, only two characterized sesquarTPSs are documented in TPS db.

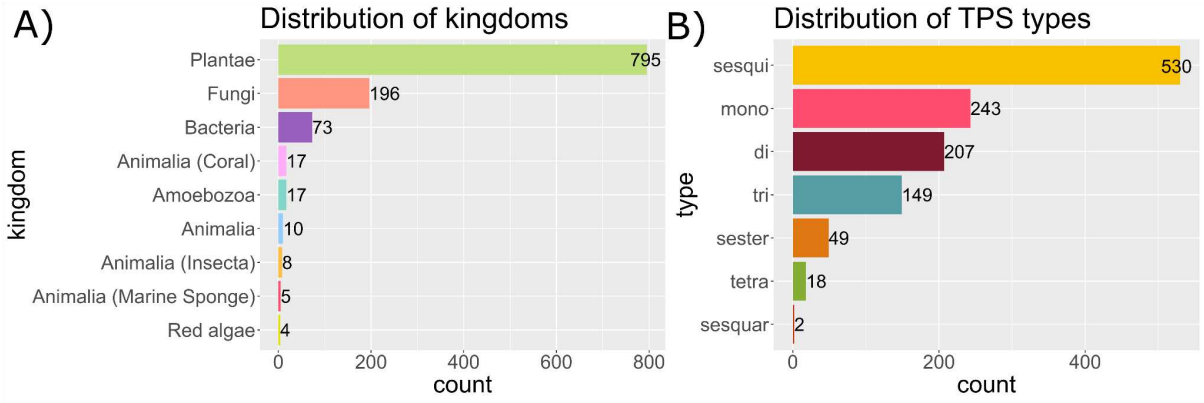

Figure 9. A) Number of TPSs in each taxonomic group B) Number of TPSs in each TPS type; one TPS can occur in more categories if it produces products of more categories

4.2.1 Pfam and SUPERFAMILY domains

Models from TPS Pfam db and TPS SUPERFAMILY db, described in Chapters 4.1.3.1 and 4.1.3.2, respectively, were employed to scan the TPS database using HMMER hmmscan (Sean R. Eddy 2023). This analysis revealed notable differences in occurrences of Pfam domains across various kingdoms, as outlined in Table 7. For instance, the Pfam model PF01397 exclusively captured sequences from plants, which is in agreement with the available taxonomical distribution from InterPro, as illustrated in Figure 5. Surprisingly, PF19086 demonstrated the ability to capture microbial-type TPSs but also a high number of plant TPSs. An unsurprising outcome was that Pfam domains failed to detect insect TPSs, as they differ from typical TPSs. Only 15 sequences lacked Pfam hits, out of which eight were TPSs from insects.

Table 7. Percentage of TPSs from TPS db containing Pfam model hits across different kingdoms. The last row provides the total percentage of sequences across all kingdoms.

	PF01397	PF03936	PF19086	PF06330	PF13243	PF13249	PF00494
Plantae	83,1	83,7	79,1	0,6	18,9	15,6	2,4
Red algae	0,0	0,0	100,0	0,0	0,0	0,0	0,0
Fungi	0.0	54,6	87,2	18,9	8,7	7,1	4,1
Bacteria	0.0	42,7	66,7	10,7	13,3	13,3	16,0
Cyanobacteria	0,0	0,0	0,0	0,0	100,0	100,0	0,0
Coral	0,0	17,6	100,0	0,0	0,0	0,0	0,0
Amoebozoa	0,0	35,3	88,2	0,0	5,9	5,9	5,9
Insecta	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Animalia	0,0	0,0	46,7	0,0	26,7	26,7	20,0
Total	58,7	72,2	79,2	4,4	16,3	13,7	3,8

An overview of the occurrences of individual SUPERFAMILY domains from TPS SUPERFAMILY db across various kingdoms is presented in Table 8. This analysis revealed results that are consistent with the distinct philosophy of the SUPERFAMILY database compared to Pfam. In the SUPERFAMILY database, models can overlap, a feature evident for TPSs from TPS db where most sequences are captured by multiple models. Notably, in more than 80% of cases, plant sequences were detected by all models employed.

Unlike Pfam, SUPERFAMILY models were able to detect even some insect TPSs. For 169 sequences, there were no SUPERFAMILY hits. Only five TPSs did not have hits from either Pfam or SUPERFAMILY. On the other hand, 950 sequences (84%) contain hits from both Pfam and SUPERFAMILY.

Table 8. Percentage of TPSs from TPS db containing SUPERFAMILYmodel hits across different kingdoms. The last row provides the total percentage of sequences across all kingdoms.

	41184	53354	53355	48261	48806	46340	47573
Plantae	83,1	83,1	84,5	84,5	84,4	84,5	84,4
Red algae	0,0	0,0	100,0	100,0	100,0	100,0	100,0
Fungi	0.0	0.0	88,8	89,8	90,3	90,3	89,8
Bacteria	0.0	0.0	68,0	70,7	68,0	70,7	68,0
Cyanobacteria	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Coral	0,0	0,0	100,0	100,0	100,0	100,0	100,0
Amoebozoa	0,0	0,0	88,2	88,2	88,2	88,2	88,2
Insecta	0.0	0.0	12,5	12,5	50	87,5	25
Animalia	0,0	0,0	46,7	46,7	46,7	46,7	46,7
Total	58,7	58,7	83,4	83,7	83,8	84,4	83,6

4.2.2 Length distribution analysis

The amino acid sequence length of each terpene synthase within the TPS db was calculated to analyze the distribution across different TPS types and kingdoms. Additionally, the lengths of sequences captured by each Pfam or SUPERFAMILY domain were compared to find any patterns.

Overall, three main peaks emerged in the length distribution of TPSs, centered around 350, 565, and 760 amino acids, as depicted in Figure 10A. However, further analysis revealed characteristic differences in length distribution among various TPS types and kingdoms, as illustrated in Figure 10B.

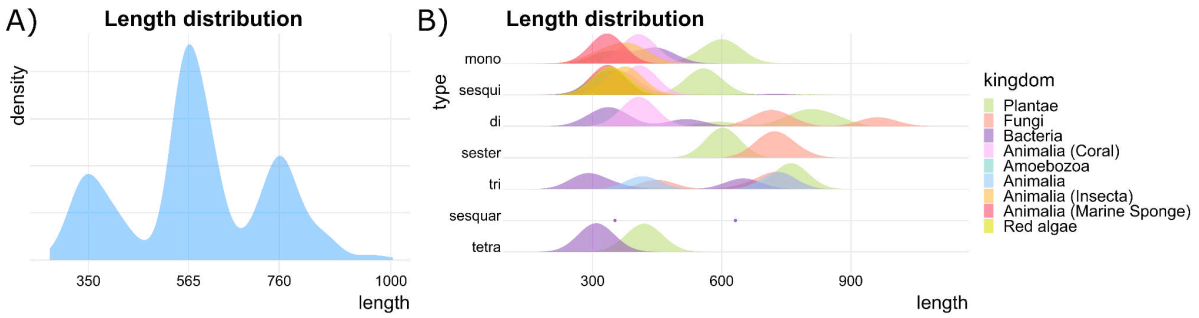

Figure 10. Length distribution of TPSs. (A) Length density plot depicting the lengths of TPSs in TPS db forming three peaks at 350, 565, and 760 amino acids. (B) Density ridge plot illustrating the length diversity of TPSs categorized by type and kingdom.

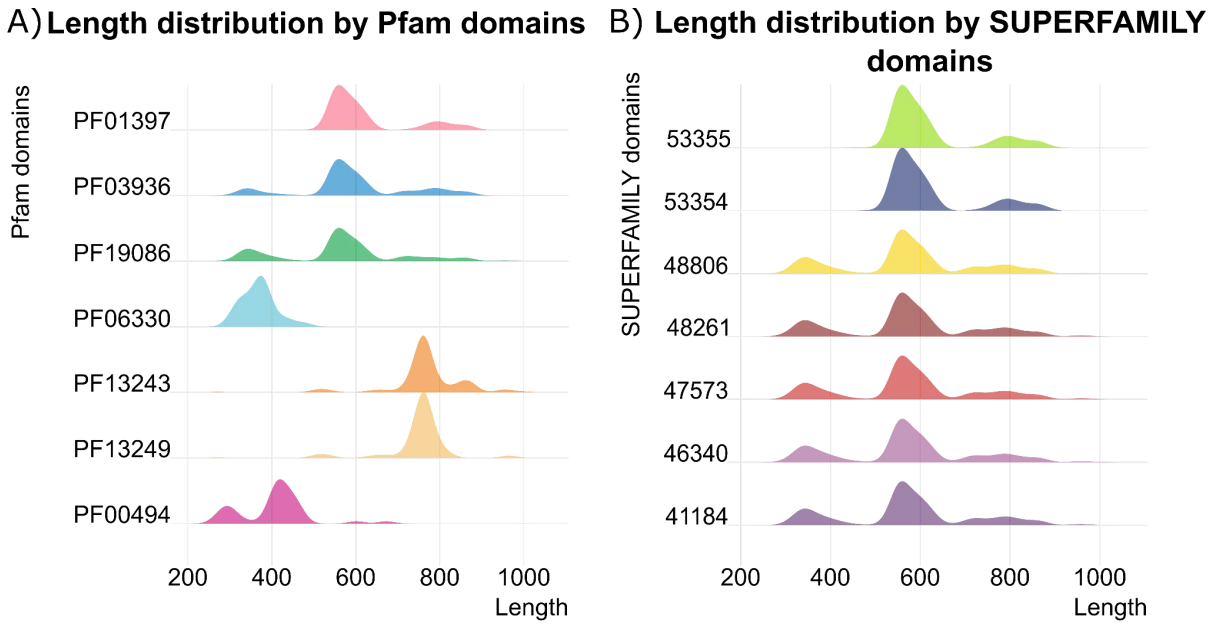

Figure 11. *Length distribution of sequences from TPS db captured by different Pfam* (A) *and SUPERFAMILY* (B) *models.*

Figure 11A demonstrates how different Pfam models capture sequences of varying lengths. This does not hold for SUPERFAMILY models because of the previously mentioned ability to capture the majority of all sequences, therefore resulting in a length distribution closely resembling the mean length, as depicted in Figure 11B.

4.2.3 Domain architecture analysis

The common combinations of Pfam and SUPERFAMILY domains in the data were analyzed.

Figure 10B illustrates the varying lengths observed for different TPS types across kingdoms. These peaks correspond to different Pfam domain architectures, as depicted in Figure 12. The most common Pfam architecture was found to be a combination of PF01397 and PF03936 domains, a typical domain architecture for plant TPSs. In plant diTPSs, this architecture can be extended by fragments of TPS Pfam domains from one or both sides, resulting in longer sequences. The second most common architecture consists of a single PF19086 domain, which is the architecture observed in mono-, sesqui-, and diTPSs from all other kingdoms, except diTPSs from Fungi, where this domain occurs duplicated. This architecture is also typical for fungal sesterTPSs. The third most common architecture, a combination of PF13249 and PF13243 domains, is typical for squalene-hopene cyclases (triTPSs). Other architectures include a single PF00494 domain of squalene or phytoene synthases, or a single PF06330 domain of trichodiene synthases (sesquiTPSs) in Fungi.

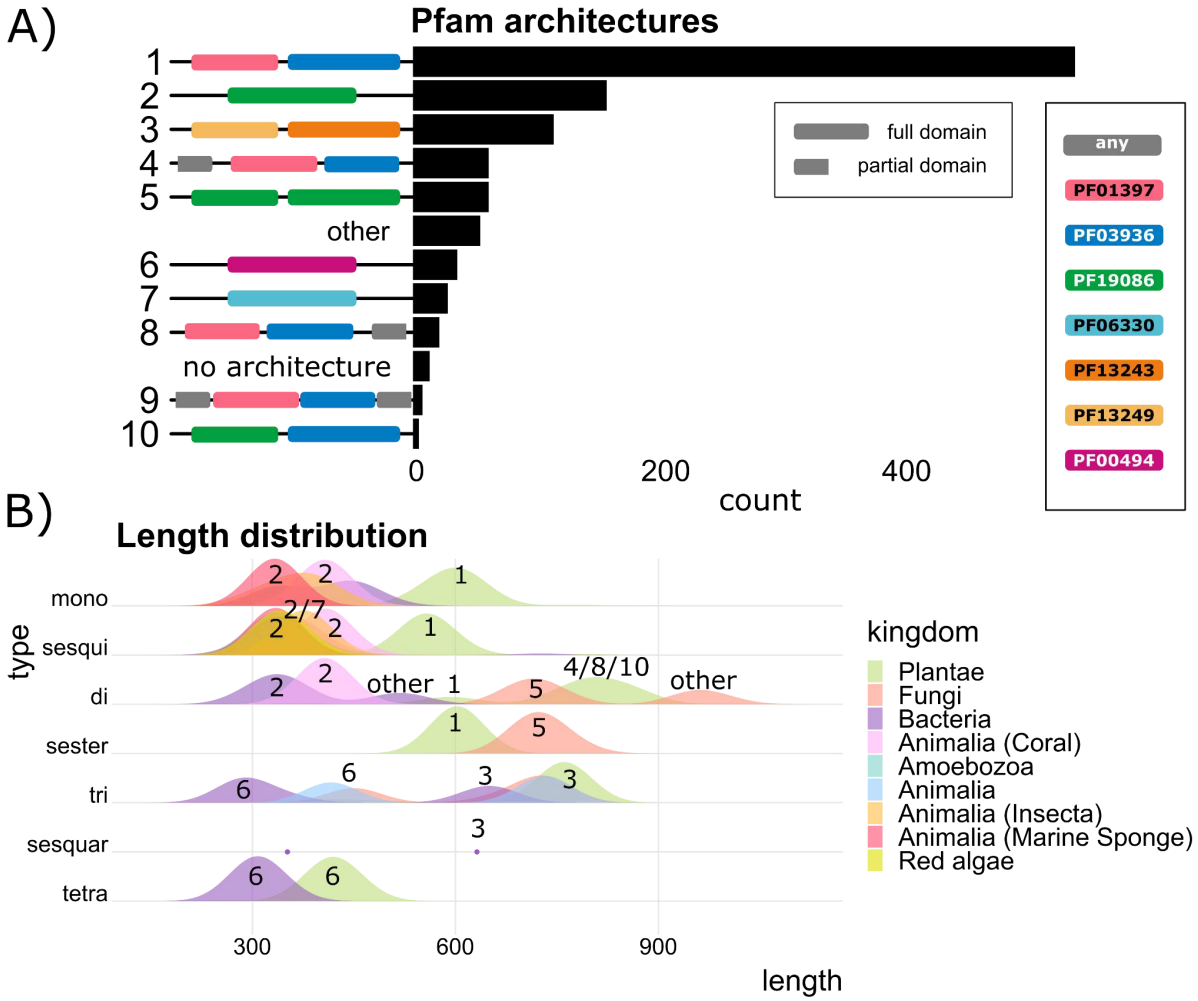

Figure 12. Analysis of Pfam domain architectures and length diversity of TPSs in TPS db. (A) A histogram depicts the observed Pfam architectures. Each architecture is schematically illustrated (actual lengths of sequences and domains are not reflected) and assigned a number. In architectures 4, 8, and 9, the grey partial domain represents any partial domain from TPS Pfam db. "Other" encompasses all remaining architectures, and "no architecture" encompasses TPSs with no Pfam hits. (B) Density ridge plot from Figure 10B illustrating the length diversity of TPSs categorized by type and kingdom, with numbers above the peaks denoting the most common architecture(s) for sequences of the corresponding kingdom and type.

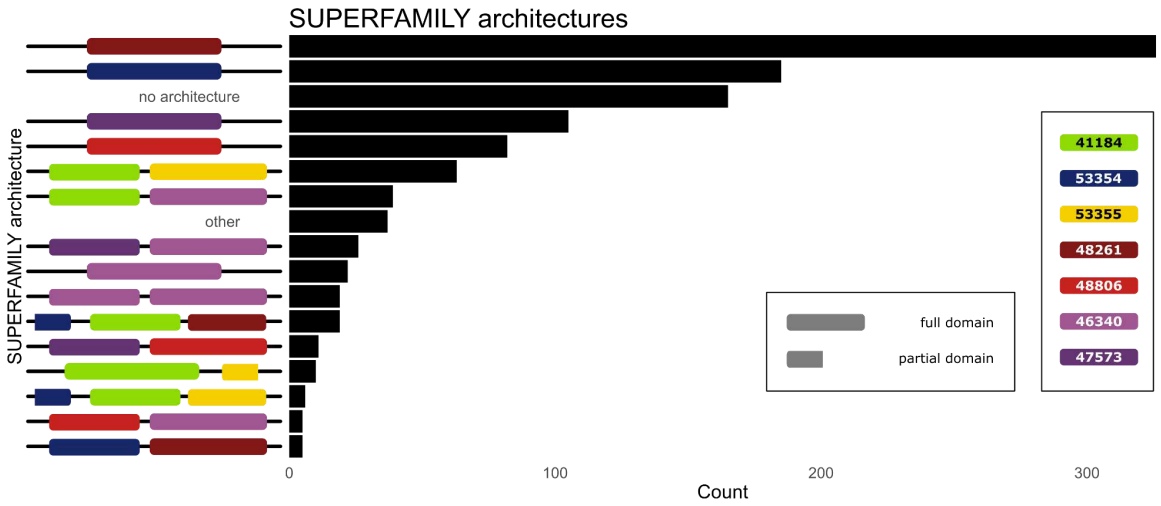

Figure 13. *A histogram depicts the observed SUPERFAMILY architectures in TPS db. Each architecture is schematically illustrated (actual lengths of sequences and domains are not reflected). "Other" encompasses all remaining architectures, and "no architecture" encompasses TPSs with no SUPERFAMILY hits.*

Figure 13 provides an overview of SUPERFAMILY domain architectures. This analysis revealed a major occurrence of single domains. The most common single-domain architectures include domains 48261, 53354, 47573, and 48806. However, domain 41184 often occurs in combination with other domains. Domain 46340 was observed both independently and in combination with other domains (including itself).

Overall, characterized TPSs from TPS db contain full domains or at least one full domain. Only five TPSs contained Pfam domain architecture consisting of a single partial domain, and only seven other TPSs contained SUPERFAMILY domain architecture consisting of a single partial domain.

4.2.4 Conserved motifs

The TPS database was analyzed to see whether the sequences contain the conserved functional motifs of terpene synthases - namely, the DDXXD and NSE/DTE motifs of Class I TPSs and the DXDD motif of Class II TPSs. These motifs were identified using the EMBOSS fuzzpro tool (Rice, Longden, and Bleasby 2000) designed for pattern searches

within protein sequences. Specifically, the presence of these motifs was determined based on predefined patterns: for the DDXXD motif, at least one hit of the following patterns indicated its presence: DDXXD or DDXX[DE]; for the NSE/DTE motif, at least one occurrence of the following patterns indicated its presence: [ND][DE]XX[ST]XX[NKR][DE], [ND]D[LIV]X[ST]XXXE, or [ND]DXX[ST]XXXE; and for the DXDD motif, an exact match was considered, i.e., DXDD.

The results of the presence of individual motifs are summarized in Table 9. Almost all mono-, sesqui-, di-, and sesterTPSs contain at least one motif. In contrast, tri-, sesquar-, and tetraTPSs possess variations of the motifs or different motifs, as the inspected motifs were present in less than 50% of the TPSs.

Among the Class I motifs, DDXXD was more frequently present than the NSE/DTE motif, possibly due to slight variations in the latter. The Class II motif DXDD was not prominently represented in the dataset.

Table 9. Frequency of conserved motifs within different TPS types.

	MonoTPS	SesquiTPS	DiTPS	SesterTPS	TriTPS	SesquarTPS	TetraTPS
At least 1 motif	99,6%	99,4%	99,5%	98,0%	18,1%	50,0%	38,9%
DDXXD (Class I)	97%	89%	79%	96%	12%	0%	39%
NSE/DTE (Class I)	64%	76%	63%	94%	2%	0%	0%
DXDD (Class II)	7%	12%	39%	12%	9%	50%	6%

4.2.5 Sequence similarity of terpene synthases and IDSs

The objective of this analysis was to investigate whether the removal of sequences highly similar to IDSs would accidentally remove terpene synthases as well.

IDS sequences were obtained from the unfiltered TPS database and divided into two classes: monofunctional IDSs and bifunctional IDS-TPSs (PTTSs), denoting sequences exhibiting both IDS and TPS activity.

The pairwise sequence identities were obtained from all-vs-all BLAST.

There was only one case of sequence identity between monofunctional IDS and TPS higher than 80%, observed for sesterTPS A0A0F4GLU2, annotated in UniProt as Geranylgeranyl pyrophosphate synthase like protein, suggesting this sequence might actually represent bifunctional IDS-TPS. However, in most cases, the similarity remained below 50%, with an average identity of 31%, as depicted in Figure 14A.

Even among bifunctional IDS-TPSs, high similarity to monofunctional IDSs was rare, with only two sequences exhibiting identity over 80% and an average identity of 32%, as depicted in Figure 14B.

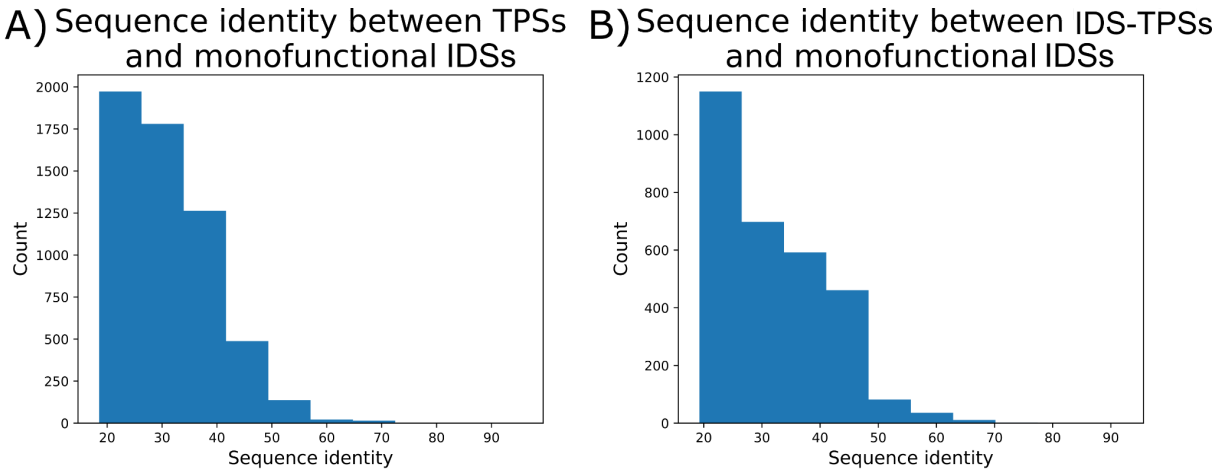

Figure 14. Histograms of sequence identity between monofunctional IDSs and TPSs (A) and bifunctional IDS-TPSs (B).

Therefore, removing sequences with more than 80% identity to monofunctional IDSs should not remove TPSs but rather IDSs that were incorporated by accident.

4.2.6 Protein sequence embeddings for TPS comparison

As previously mentioned, protein sequence embeddings are an alternative representation of protein sequences that are useful for sequence comparison. Embeddings were generated using the ESM2 model esm2_t6_8M_UR50D, which comprises 6 layers, 8 million parameters, and yields embeddings of dimension 320 (Lin et al. 2023). This model was chosen as ESM2 models are among the most popular protein language models, and this version has the smallest size and relatively low resource requirements. The objective was to test if this small model could produce meaningful

representations of TPSs as it could be potentially used on a large number of mined sequences to compare them. In Figure 15, the projections of the embeddings using t-SNE are illustrated. The Figure shows that the embedded TPSs exhibit clustering based on both the type and kingdom, highlighting that sequence embeddings may provide a useful representation for comparing TPSs.

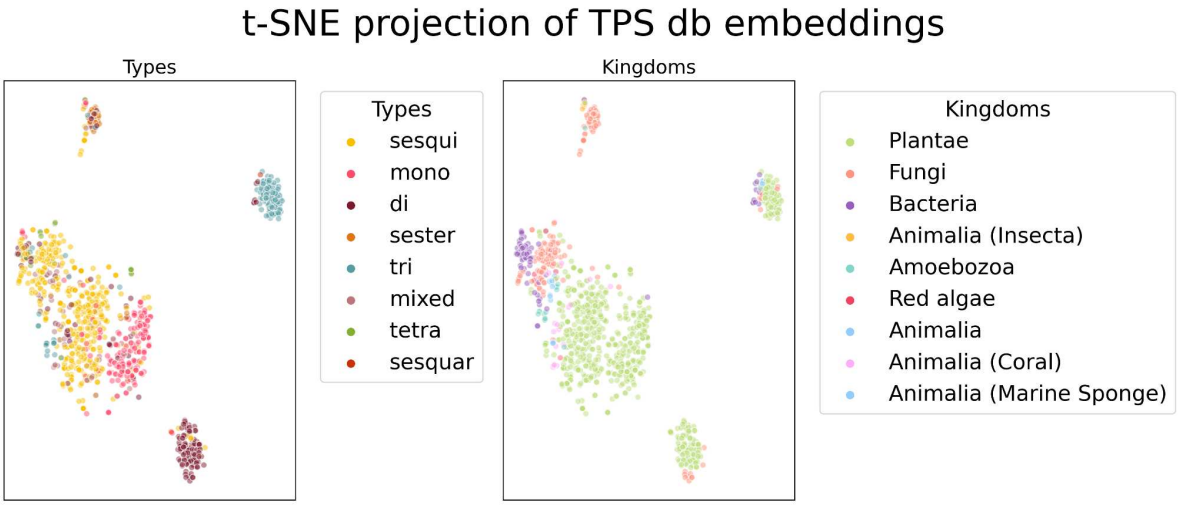

Figure 15. t-SNE projection of protein embeddings of TPSs from TPS db colored according to their type (left) and kingdom (right). Points cluster according to both type and kingdom.

4.3 Mining of putative terpene synthases

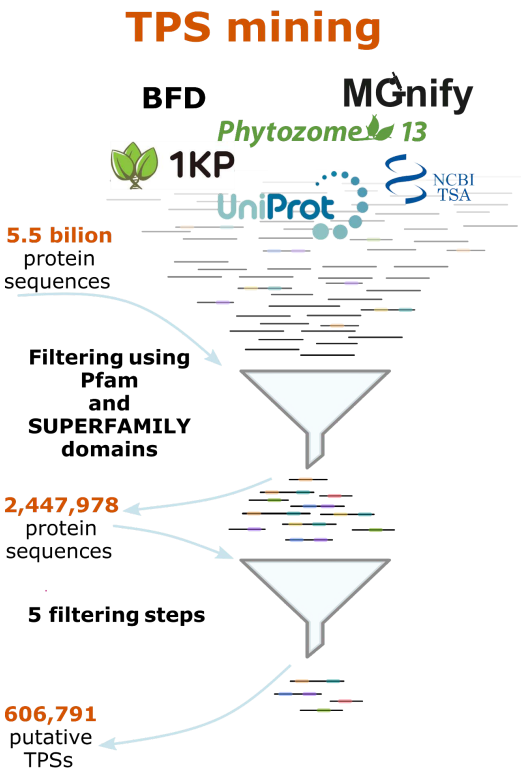

Figure 16. Schematic diagram of the mining process.

The mining process was executed on MetaCentrum[4] through a custom Snakemake pipeline (Köster and Rahmann 2012) developed for this project (see attachment A.4). Programs required for the whole process were installed using Anaconda (Anaconda Inc. 2020) and pip, and their versions are listed in the GitHub repository (see attachment A.4).

To enable parallel processing, each sequence database fasta file was split using the fasta-splitter tool (Kryukov 2021) into smaller chunks of 1,000,000 sequences. Each file was first length-filtered to ensure the length of each sequence between 20 and 100,000 amino acids, addressing issues with very short and very long sequences in subsequent stages.

[4] MetaCentrum, organization of the Czech National Grid Organization providing distributed computing infrastructure
https://metavo.metacentrum.cz

For Pfam mining, TPS Pfam db models (see Chapter 4.1.3.1) were downloaded from the InterPro website (Paysan-Lafosse et al. 2023). They were merged into a database, which was then compressed and indexed using the hmmpress command from HMMER (Sean R. Eddy 2023).

Similarly, HMM models from the TPS SUPERFAMILY db (see Chapter 4.1.3.2) were downloaded from the SUPERFAMILY database version 1.75 and merged into a database, which was compressed and indexed with hmmpress.

The actual mining procedure utilized the hmmsearch command from HMMER (Sean R. Eddy 2023), with input consisting of the fasta file chunks, the TPS Pfam db, and TPS SUPERFAMILY db. Any sequence identified by any model from TPS Pfam db or TPS SUPERFAMILY db was mined.

The only special approach during the mining was used for the BFD database as it contained already preclustered sequences. Taking advantage of the preclustering, the mining process here took two steps. Each cluster contained a representative sequence, and in the first step, mining was done only on the representative sequences. In the second step, all cluster sequences were used only if the representative sequence had any Pfam or SUPERFAMILY hit.

4.4 Enhancing the reliability through sequence filtering

To enhance the reliability of the mined sequences as potential terpene synthases with the capability to produce terpenes, a series of filtering steps were implemented based on insights from the literature and analysis of characterized terpene synthases described in Chapter 4.2. The filtering process comprised five steps denoted as: length filtering, stronger domain hit from other families, presence of a functional motif, bad domain architecture, and possible IDS.

According to literature and the analysis of characterized terpene synthases from TPS db, terpene synthases typically exhibit lengths between 300 and 1100 amino acids. Sequences falling outside this range were filtered out.

To further minimize the risk of a false hit, sequences with a stronger hit to any Pfam or SUPERFAMILY domain not associated with terpene synthases were filtered out. For the Pfam database, except from domains in TPS Pfam db, the following domains were considered as associated with terpene synthases: PF00348 (Polyprenyl synthetase), PF00432 (Prenyltransferase and squalene oxidase repeat), and PF02458 (Transferase family). For SUPERFAMILY, any domain in the 48576 and 48239 superfamilies was considered.

Class I terpene synthases typically contain the DDXXD and NSE/DTE motifs. Class II terpene synthases typically contain the DXDD motif. The motifs were scanned using EMBOSS fuzzpro as described in 2.4.2. Sequences that did not contain at least one motif were filtered out.

Within TPS db, only five sequences (<1%) fall into the following Pfam and SUPERFAMILY "bad domain architectures": (i) absence of both Pfam and Supfam architectures, (ii) absence of Pfam architecture but presence of a single partial Supfam domain, (iii) absence of Supfam architecture but presence of a single partial Pfam domain, or (iv) presence of both a single partial Pfam domain and a single partial Supfam domain architectures. This means that functional terpene synthases almost never consist of just a single partial domain architecture. Thus, all mined sequences falling into those four categories were filtered out since there is a risk that these sequences would not be functional.

Lastly, sequences highly similar to IDSs were removed to mitigate the risk of accidentally mining these enzymes. Monofunctional IDSs from the TPS db, as described in Chapter 4.2.5, were BLASTed with the mined sequences, and all sequences with a sequence identity above 80% were filtered out.

These sequences are further referred to as candidate terpene synthases, TPS candidates or putative TPSs.

4.5 Sequence annotation

Computational and manual approaches were combined to annotate the candidate terpene synthase sequences. The annotation process involved gathering information about the source organism and its NCBI taxon ID. In cases when it was possible to acquire the taxon ID, the taxoniq Python library ("Taxoniq: Taxon Information Query - Fast, Offline Querying of NCBI Taxonomy and Related Data," n.d.) was utilized to fetch the species lineage, extracting details such as superkingdom, kingdom, and phylum.

When enough information was available, metagenomic sequences were placed in one of the following categories as per NCBI taxonomy classification: engineered, environmental, host-associated, mixed, and unknown.

Since sequences in the UniParc database represent sequences from multiple databases, they were mapped back to the original entries using the UniProt API to get the available annotations.

In addition, the annotation also integrated the Pfam and SUPERFAMILY domain architectures, along with boolean indicators of hit of each TPS Pfam db and TPS SUPERFAMILY db domains.

Lastly, the ID of the closest characterized TPS from TPS db was added based on sequence similarity and the Euclidean distance of the protein embeddings (see attachment A.3).

4.6 Scoring the terpene synthase candidates

The objective of this section was to show how resulting sequences could be prioritized for functional characterization. The aim would be to identify candidate sequences that have a higher chance to be functional terpene synthases but also have a higher chance to yield novel terpene products. For this purpose, two separate scores were proposed: the reliability score and the novelty score. The scores can be weighted and combined, allowing a flexible approach to prioritize one criterion over the other. It is essential to

note that the proposed scores are simplistic and serve as guiding tools rather than definitive selection criteria.

4.6.1 Reliability score

The reliability score aims to capture if a candidate sequence has properties observed in the characterized sequences and the certainty of being a functional terpene synthase. The score is derived by summing up six partial scores, each ranging from 0 to 1.

These six partial scores include the methionine score (S_{met}), observed Pfam domain architecture score (S_{p_arch}), observed SUPERFAMILY domain architecture score (S_{s_arch}), strongest Pfam hit c-Evalue score (S_{p_evalue}), strongest SUPERFAMILY hit c-Evalue score (S_{s_evalue}), and the presence of domain hit from both TPS Pfam db and TPS SUPERFAMILY db score ($S_{p\&s}$).

$$S_{reliability} = S_{met} + S_{p\_arch} + S_{s\_arch} + S_{p\_evalue} + S_{s\_evalue} + S_{p\&s}$$

Figure 17. The formula to calculate the reliability score.

The methionine score (S_{met}) is a binary score of 1 for all sequences starting with methionine and 0 otherwise. Functional terpene synthases start with methionine.

$$S_{met} = \begin{cases} 1, & \text{if sequence starts with methionine} \\ 0, & \text{otherwise} \end{cases}$$

Figure 18. The formula to calculate the methionine score.

The observed Pfam architecture score (S_{p_arch}) and observed SUPERFAMILY architecture score (S_{s_arch}) are binary scores of 1 if a candidate sequence consists of Pfam domain architecture observed in the TPS db (or SUPERFAMILY domain architecture, respectively) and 0 otherwise.

$$S_{p\_arch} = \begin{cases} 1, & \text{if sequence has observed Pfam architecture} \\ 0, & \text{otherwise} \end{cases}$$
$$S_{s\_arch} = \begin{cases} 1, & \text{if sequence has observed SUPERFAMILY architecture} \\ 0, & \text{otherwise} \end{cases}$$

Figure 19. The formula to calculate the observed Pfam architecture score and observed SUPERFAMILY architecture score.

The following partial scores reflect the strength of a Pfam domain hit (S_{p_evalue}) or SUPERFAMILY domain hit (S_{s_evalue}), respectively. The strength of a hit can be expressed by the conditional E-value (c-Evalue) of the domain hit provided by HMMER. The smaller the c-Evalue, the stronger the hit is. The assumption was that if the candidate sequence has a strong domain hit, it is more reliable. For each candidate sequence, the lowest c-Evalue of all TPS Pfam domains (or TPS SUPERFAMILY domains, respectively) was used. Since there was a large range of exponential values, the values were transformed by taking the negative logarithm of the c-Evalues and then normalized to the range from 0 to 1. After the transformation, the strongest hits got a value of 1, and the weakest hits got a value of 0.

$$S_{p\_evalue} = normalize(-log(c\text{-}Evalue_{min(p)}))$$
$$S_{s\_evalue} = normalize(-log(c\text{-}Evalue_{min(s)}))$$

Figure 20. The formula to calculate the strongest Pfam hit c-Evalue score and strongest SUPERFAMILY hit c-Evalue score (not rigorously formalized).

Finally, the last score ($S_{p\&s}$) checks if the candidate sequence carries at least one domain hit from both Pfam and SUPERFAMILY. The assumption here was that the presence of domain hits from two domain databases enhances the reliability of the candidate sequence as the majority (84%) of sequences from TPS db contain both Pfam and SUPERFAMILY hits.

$$S_{p\&s} = \begin{cases} 1, & \text{if sequence has hit from both Pfam and SUPERFAMILY} \\ 0, & \text{otherwise} \end{cases}$$

Figure 21. The formula to calculate the presence of domain hit from both TPS Pfam db and TPS SUPERFAMILY db score.

4.6.2 Novelty score

The aim of the novelty score is to identify candidate sequences with a higher chance of producing novel terpene products. Consisting of four subscores, each ranging from 0 to 1, the novelty score seeks to penalize candidate sequences close to the characterized sequences while prioritizing those that are more distant. The four approaches employed for each subscore are taxonomical categorization, sequence similarity network, phylogenetic tree, and protein embedding distance, resulting in taxonomic score (S_{tax}), SSN score (S_{ssn}), phylogenetic score (S_{phylo}), and embedding product score ($S_{product}$).

$$S_{novelty} = S_{tax} + S_{ssn} + S_{phylo} + S_{product}$$

Figure 22. The formula to calculate the novelty score.

The novelty score was then tested on a subset of the characterized sequences from TPS db and is discussed in Chapter 6.

4.6.2.1 Taxonomic score

The assumption is that there is a higher chance of discovering a terpene synthase that yields a novel terpene when characterized sequences from the same organism or higher taxonomic rank are absent. Recognized taxonomic ranks used in the score include phylum, class, order, family, genus, and species. If there are no characterized sequences in a particular taxonomic rank, it is denoted as uncharacterized. The score ranges from 1 for an uncharacterized phylum to 0 for a characterized species, missing taxonomic information, or sequences from metagenomic samples. The scoring system assigns the score as follows: 1 is assigned if the candidate sequence originates from an

uncharacterized phylum, 0.8 from an uncharacterized class, 0.6 from an uncharacterized order, 0.4 from an uncharacterized family, 0.2 from an uncharacterized genus, 0.1 if uncharacterized species and 0 from a characterized species, cases when the taxonomic information is missing or the candidate sequence originates from a metagenomic sample.

$$S_{tax}(t) = \begin{cases} 1, & \text{t in uncharacterized phylum} \\ 0.8, & \text{t in uncharacterized class} \\ 0.6, & \text{t in uncharacterized order} \\ 0.4, & \text{t in uncharacterized family} \\ 0.2, & \text{t in uncharacterized genus} \\ 0.1, & \text{t in uncharacterized species} \\ 0, & \text{t in characterized species/metagenomic/missing} \end{cases}$$

Figure 23. The formula to calculate the taxonomic score.

4.6.2.2 SSN score

For the purpose of visualizing the result of the mining but also prioritization of the candidates, a sequence similarity network was created and leveraged. The SSN was constructed following the guidelines outlined in the supplementary material of (Copp et al. 2018). The SSN was created by reducing the dataset of TPS candidates and performing an all vs. all blast search.

The reduced dataset was created by similarity clustering using a threshold of 50% with CD-hit (W. Li and Godzik 2006; Fu et al. 2012). This threshold was selected to obtain a manageable number of cluster representatives for downstream steps, including visualization in Cytoscape. A higher threshold could be used in a scenario with fewer mined sequences. Lastly, all characterized terpene sequences were also added to the dataset.

The sequence similarity network from this reduced dataset was created by using the result of an all-vs-all blast search. Edges were formed between sequences, represented as nodes, with blast e-value smaller than 10e-50. This threshold was again chosen to maintain a manageable number of edges.

Clusters within the network can be visually observed, and their characteristics become apparent when annotations such as terpene product type, superkingdom, or the presence of individual domains are visualized. The Louvain method from the Networkx python package (Hagberg, Swart, and Schult 2008) was employed to automate the identification of the clusters. This method was developed to detect communities in large networks. In this case, we can consider clusters of sequences with similar features as a community. The method uses only the structure of the network to identify the clusters. The clusters were then analyzed to see how many characterized sequences they contain and how many candidate sequences they actually represent (since in the network, only sequence similarity representatives are used from the candidate sequences).

The underlying assumption when creating the SSN score (S_{ssn}) was that large clusters with no or few characterized sequences could be more interesting for exploration. The SSN score tries to capture both the degree of exploration within the cluster, using the percentage of uncharacterized sequences in the total number of sequences, and also the cluster size, using a normalized score of the logarithm of the total size.

$$S_{ssn}(t) = \frac{\text{percentage of uncharacterized}(cluster_t) + normalize(log(size(cluster_t)))}{2}$$

Figure 24. The formula to calculate the SSN score (not rigorously formalized).

4.6.2.3 Phylogenetic score

A phylogenetic tree was constructed from the reduced dataset (see Chapter 4.6.2.2) and followed a standard procedure consisting of computing an MSA (using Clustal omega (Sievers et al. 2011)), MSA trimming (using TrimAl (Capella-Gutiérrez, Silla-Martínez, and Gabaldón 2009)) and finally, tree construction (using FastTree2 (Price, Dehal, and Arkin 2010)). FastTree2 is a popular method for large datasets as it is fast and uses less memory. The method starts by creating an initial topology using the Neighbor-Joining algorithm, which is then refined using nearest-neighbor rearrangements and subtree-pruning-regrafting, and finally, it is optimized by the Maximum Likelihood algorithm.

The assumption in the phylogenetic score (S_{phylo}) is that there could be a higher chance of novelty discovery in a clade that is large and consists only of uncharacterized sequences. For each candidate sequence, the size (number of leaves) of the largest clade it belongs to, containing only uncharacterized sequences, was calculated. These numbers were then transformed by taking a logarithm and normalized between 0 and 1. For the sequences not represented in the tree, the score of their clustering representative was used.

$$S_{phylo}(t) = normalize(log(\text{size of largest clade containing t and only uncharacterized sequences}))$$

Figure 25. The formula to calculate the phylogenetic tree score (not rigorously formalized).

4.6.2.4 Embedding product score

The embedding product score ($S_{product}$) aims to compare candidate terpene synthases to characterized terpene synthases and reflect how interesting their products are.

In the TPS db, all possible products were selected, and their number of occurrences was counted. For each sequence, only the product with the smallest number of occurrences and this occurrence count were kept. The number of occurrences serves as a measure of how interesting the product is. This number was then normalized between 0 and 1 and subtracted from 1.

The protein embedding distance was employed in this score. The Euclidean distances between the embeddings of uncharacterized candidate sequences and characterized sequences were calculated. This process enabled the assignment of the closest characterized sequence to each uncharacterized sequence and the usage of its product score.

$$S_{product}(t) = product\ score(closest\ characterized(t))$$

$$product\ score(t) = 1 - normalize(\#\ occurences(rarest\ product\ \ in\ TPS\ db(t)))$$

Figure 26. The formula to calculate the embedding product score (not rigorously formalized).

5 Results

5.1 Overview of the TPS mining

Sequence mining using Pfam domains resulted in 1,456,732 TPS-like sequences, and using SUPERFAMILY domains resulted in 1,312,531 sequences, together 2,447,978 unique sequences. Clustering with 90% identity resulted in 1,207,771 clusters. Surprisingly, there was not a big overlap in the sequences mined by Pfam and SUPERFAMILY domains. For comparison, in TPS db, 84% of sequences contain hits from both Pfam and SUPERFAMILY.

This dataset consisted of 48% from data from the BFD database, 30% from MGnify, 15% from UniParc, 5% from TSA, 1% from OneKP, and 1% from Phytozome. However, the dataset was further filtered to increase the quality of the final dataset, as described in Chapter 4.4. Length filtering resulted in the removal of many sequences, especially from the metagenomic databases, where more than 70% of sequences were filtered out. The most common cases were short sequences with domain hits from TPS Pfam or TPS SUPERFAMILY db, but they were likely too short to be functional.

The final dataset of TPS candidates after the five rounds of filtering comprises a total of 606,791 sequences (see attachment A.3). The majority of TPS candidates originate from UniParc, MGnify, and BFD repositories, as depicted in Figure 27.

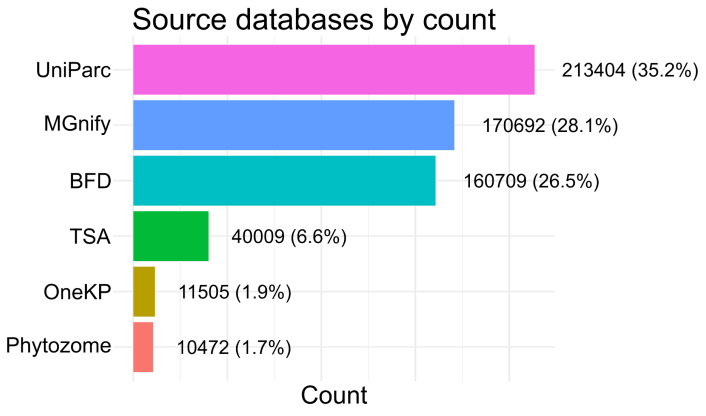

Figure 27. Number and percentage of TPS candidates from each database.

The length distribution of TPS candidate sequences is predominantly located within the range of 300 to 500 amino acids. The peaks at 565 and 760 amino acids, which were characteristic for characterized TPSs in TPS db (Figure 10A), are also apparent in this dataset (Figure 28).

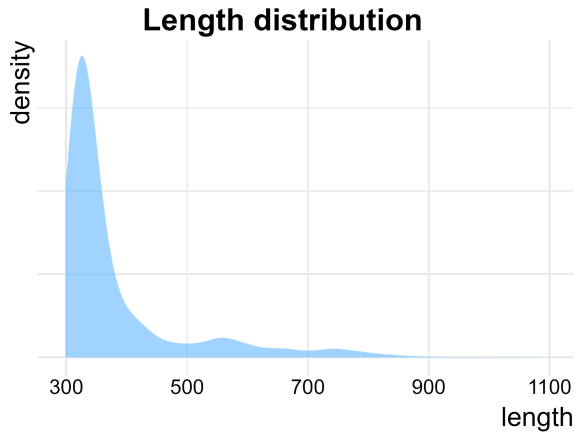

Figure 28. Length density plot depicting the lengths of TPS candidates. The density peak on the left side appears trimmed as the minimum sequence length was set to 300 amino acids.

Nearly half of the TPS candidates originate from metagenomic sources (Figure 29), underscoring the richness of metagenomes as a valuable data source for sequence mining. A closer examination of the metagenomic candidates shows that the majority originates from various environmental samples. The taxonomic analysis revealed that the largest represented superkingdom is Bacteria, constituting 25.9% of the dataset, followed by Eukarya at 18.6%. Archaea represents a small fraction of 0.9%. 84 candidates were also identified among viruses. In Bacteria, there are several notable phyla that contain TPS candidates but do not contain previously experimentally characterized TPSs. These include *Campylobacterota*, *Planctomycetota*, *Acidobacteriota*, *Verrucomicrobiota*, *Thermodesulfobacteriota*, *Spirochaetota*, *Gemmatimonadota*, and *Deinococcota*. Among the previously characterized phyla, notable taxonomic classes without previously characterized TPSs are primarily found in *Pseudomonadota*, *Bacteroidota*, and *Bacillota*.

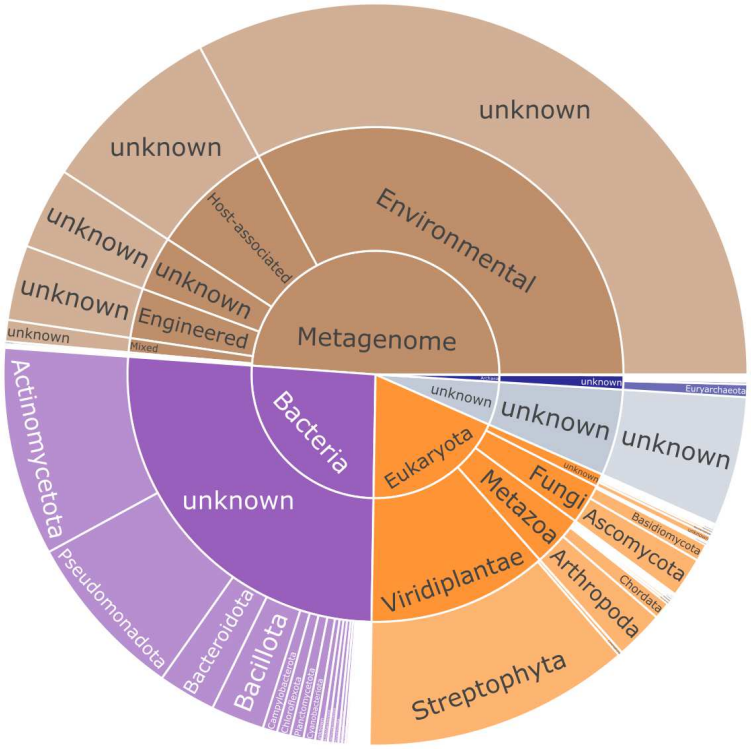

Figure 29. Sunburst plot of the taxonomy of TPS candidates (superkingdom, kingdom, phylum).

A closer look at TPS candidates from Eukaryota, illustrated in Figure 30, reveals that the majority group within Eukaryotic candidates are plants (63.2%). The second largest group comprising 18.2% are TPS candidates from Metazoa, closely followed by TPS candidates from Fungi (14.1%). Several notable eukaryotic phyla lack previously characterized TPSs, including *Bacillariophyta* (Diatoms), *Nematoda* (roundworms), *Mollusca*, *Haptophyta*, and *Mucoromycota*. Among the previously characterized phyla, notable taxonomic classes without previously characterized TPSs are primarily found in *Arthropoda*, *Chordata*, *Chlorophyta*, *Basidiomycota*, *Streptophyta*, and *Ascomycota*.

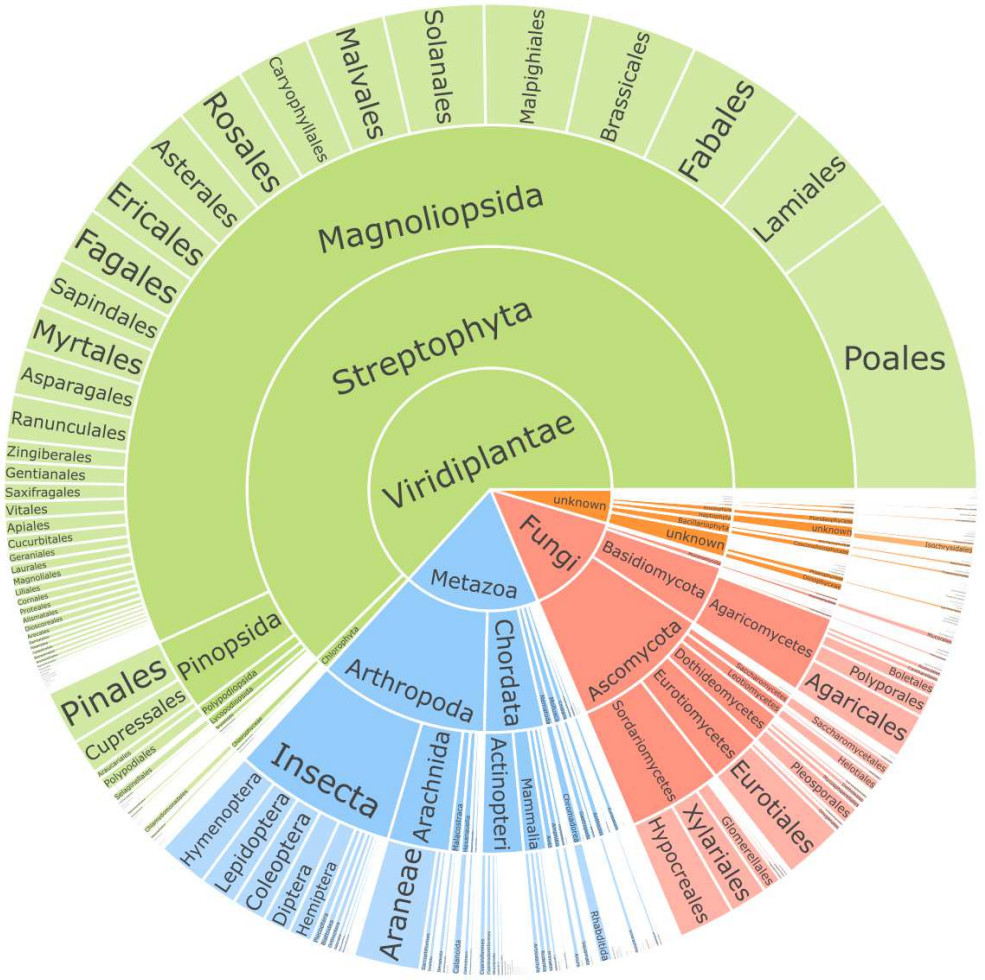

Figure 30. Sunburst plot of the taxonomy of eukaryotic TPS candidates (kingdom, phylum, class, order).

Exploring the length distribution among different superkingdoms revealed that only candidate sequences from Eukaryota exhibit a larger number of sequences that are longer than 500 amino acids. Candidate sequences from other superkingdoms exhibit lengths primarily between 300 and 400 amino acids, as depicted in Figure 31.

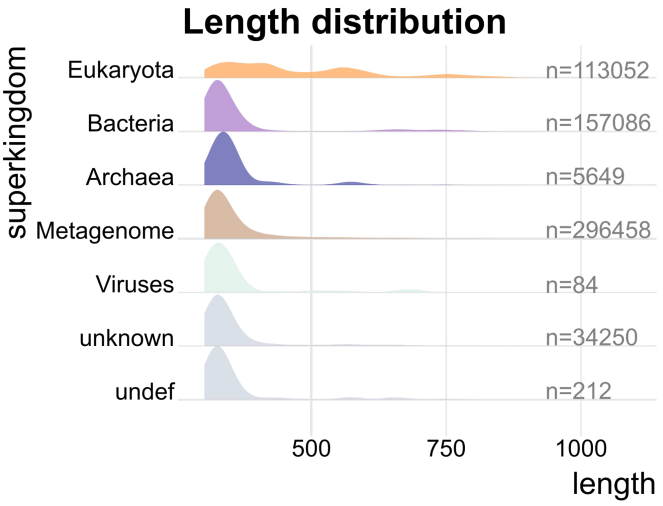

Figure 31. Density ridge plot illustrating the length diversity of TPS candidates categorized by superkingdom.

Eukaryotic sequences were thus further analyzed, focusing on Plants, Fungi, Metazoa, and other eukaryotic candidates without an assigned kingdom (Figure 32). From this, it can be seen that the longer sequences can be primarily attributed to plant sequences, although some longer sequences are found also in other groups.

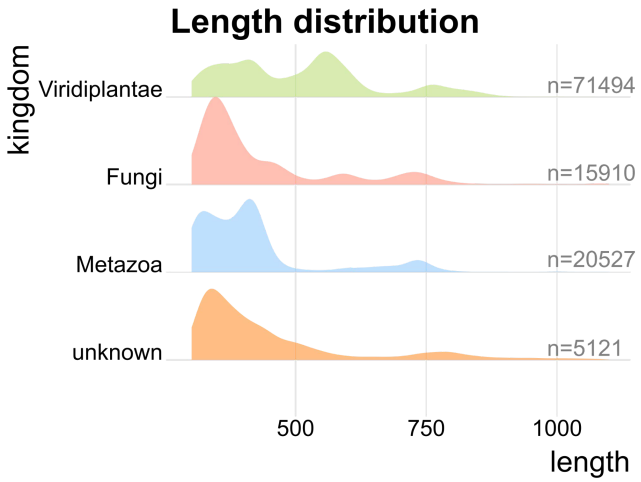

Figure 32. Density ridge plot illustrating the length diversity of eukaryotic TPS candidates categorized by the kingdom.

Out of all candidate sequences, 103,637 sequences have at least one Pfam hit. The Pfam domain that got the most hits is domain PF00494 (Squalene/phytoene synthase),

followed by domain PF19086 (Terpene synthase family 2, C-terminal metal binding). Results are summarized in Table 10.

Table 10. Percentage of TPS candidates containing Pfam model hits across different superkingdoms. The last row provides the total percentage of sequences across all superkingdoms.

	PF01397	PF03936	PF19086	PF06330	PF13243	PF13249	PF00494
Eukaryotes (all)	40,1	48,8	52,1	2,0	8,3	7,8	7,3
Eukaryotes - Plantae	63,3	69,0	65,7	0,2	8,3	7,6	6,5
Eukaryotes - Fungi	0,1	30,2	54,0	12,7	4,2	4,3	9,9
Eukaryotes - Metazoa	0,4	4,0	12,8	0,2	12,0	12,0	3,2
Eukaryotes - other	0,1	4,8	12,7	0,2	6,0	6,1	26,3
Bacteria	0,0	7,0	10,8	1,8	5,3	5,2	16,6
Archaea	0,1	3,6	5,3	0,0	2,4	2,3	28,3
Metagenome	0,2	1,8	1,7	0,1	5,0	4,9	18,7
Viruses	0,0	0,0	1,9	0,0	8,3	8,3	20,2
Total	7,7	11,7	13,6	0,9	5,8	5,7	16,1

Different Pfam models capture candidate sequences of different lengths. This is consistent with the same analysis performed for TPS db (Figure 11A); however, now there is a larger peak between 300 and 400 amino acids as these sequences are enriched in the mined dataset.

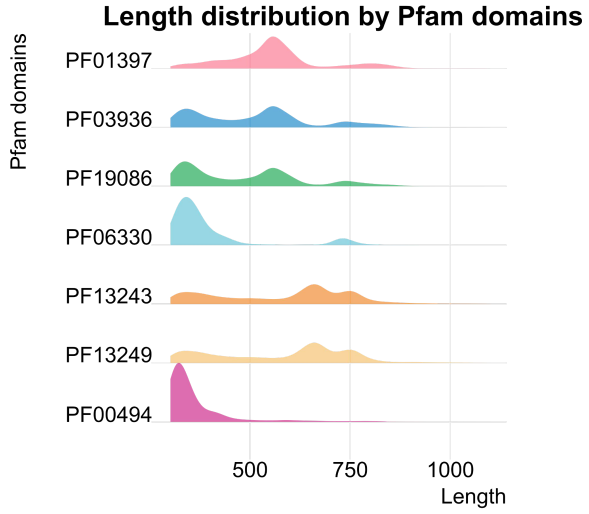

Figure 33. Length distribution of TPS candidate sequences captured by different Pfam models.

Out of all candidate sequences, 478,058 sequences have at least one SUPERFAMILY domain. Three SUPERFAMILY domains are very frequent - 48806, 46340, and 47573. Domains 48261 and 53355 are most common among plants and fungi. Domains 41184 and 53354 occur almost exclusively in plants, which was also observed for characterized sequences (Table 8).

Table 11. Percentage of TPS candidates containing SUPERFAMILY model hits across different superkingdoms. The last row provides the total percentage of sequences across all superkingdoms.

	41184	53354	53355	48261	48806	46340	47573
Eukaryotes (all)	41,2	41,2	59,7	58,8	84,0	87,0	79,7
Eukaryotes - Plantae	64,9	64,9	69,3	69,1	84,6	88,3	84,3
Eukaryotes - Fungi	0,1	0,1	68,8	68,0	86,2	87,1	80,0
Eukaryotes - Metazoa	0,4	0,4	28,7	25,8	84,2	85,7	69,6
Eukaryotes - other	0,1	0,1	20,9	19,9	68,3	73,5	55,5
Bacteria	0,0	0,0	27,7	15,7	76,0	83,8	74,2
Archaea	0,1	0,1	26,3	13,5	71,9	76,7	70,4
Metagenome	0,2	0,2	21,9	9,0	73,8	81,5	72,3
Viruses	0,0	0,0	23,8	6,0	64,3	75,0	65,5
Total	7,9	7,9	30,3	20,1	76,1	83,0	74,0

Major differences in length distribution captured by individual models are apparent only for models 41184 and 53354, which capture longer sequences (Figure 34). As mentioned earlier, these models almost exclusively capture plant sequences that tend to be longer.

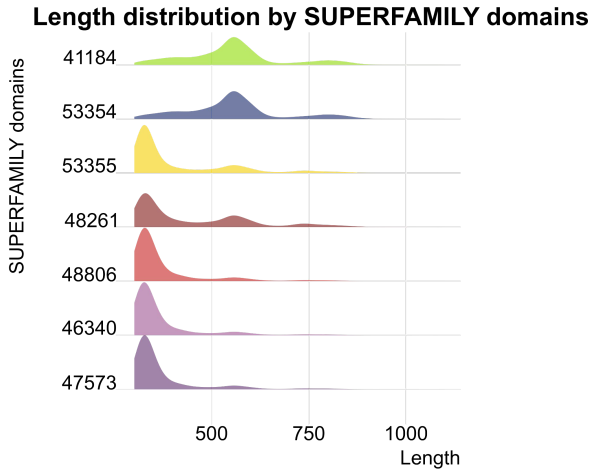

Figure 34. Length distribution of TPS candidate sequences captured by different SUPERFAMILY models.

Only 71,573 sequences have both at least one Pfam and one SUPERFAMILY domain. In contrast, in the ground-truth TPS db, 84% of sequences have both Pfam and SUPERFAMILY hits. Therefore, sequences with only Pfam or only SUPERFAMILY model hits may not represent very reliable matches. This limitation is reflected in the reliability score. However, although many sequences were captured only by SUPERFAMILY models, in most cases, multiple models captured them.

For Pfam architectures, the most common architectures are generally the ones that were also observed for characterized sequences in TPS db (Figure 12A), just in different frequencies. The most common architecture (34.6% from sequences with some Pfam architecture) is a single PF00494 domain. The second most common architecture is PF01397+PF03936 (14.8%), which is the most common architecture in the characterized TPSs. The third most common architecture is a single PF19086 domain (10%), and the fourth most common architecture is PF13249+PF13243 (7.6%) (which was the third most common architecture in the characterized TPSs).

For SUPERFAMILY, the same trend was observed as for characterized TPSs, which is that most sequences have architecture consisting of just a single domain. In the case of the candidate sequences, that was domain 46340, while for characterized TPSs, these were mostly domains 48261 and 53355.

5.2 Exploration of the TPS space using phylogenetics

To visualize the sequence space, the phylogenetic tree from Chapter 4.6.2.3 is presented here along with various annotations, including the indicators of individual Pfam and SUPERFAMILY domains along with superkingdoms/kingdoms of origin, reliability, and novelty scores (see attachment A.4). In the phylogenetic tree, only the representative sequences (from 50% sequence identity clustering) are shown. The tree was annotated and visualized in iTOL The tree is displayed, ignoring the branch lengths. It is possible to see that sequences from the same superkingdoms and with the same domain occurrences cluster together. It is also possible to see that some parts of the tree are highly characterized; however, many clades lack characterized sequences, including a large clade without characterized sequences formed mainly from metagenomic sequences (Figure 35). The visualized tree also contains the reliability score, and it is possible to observe that the parts that are mostly uncharacterized score lower, but on the other hand, they score higher in the novelty score.

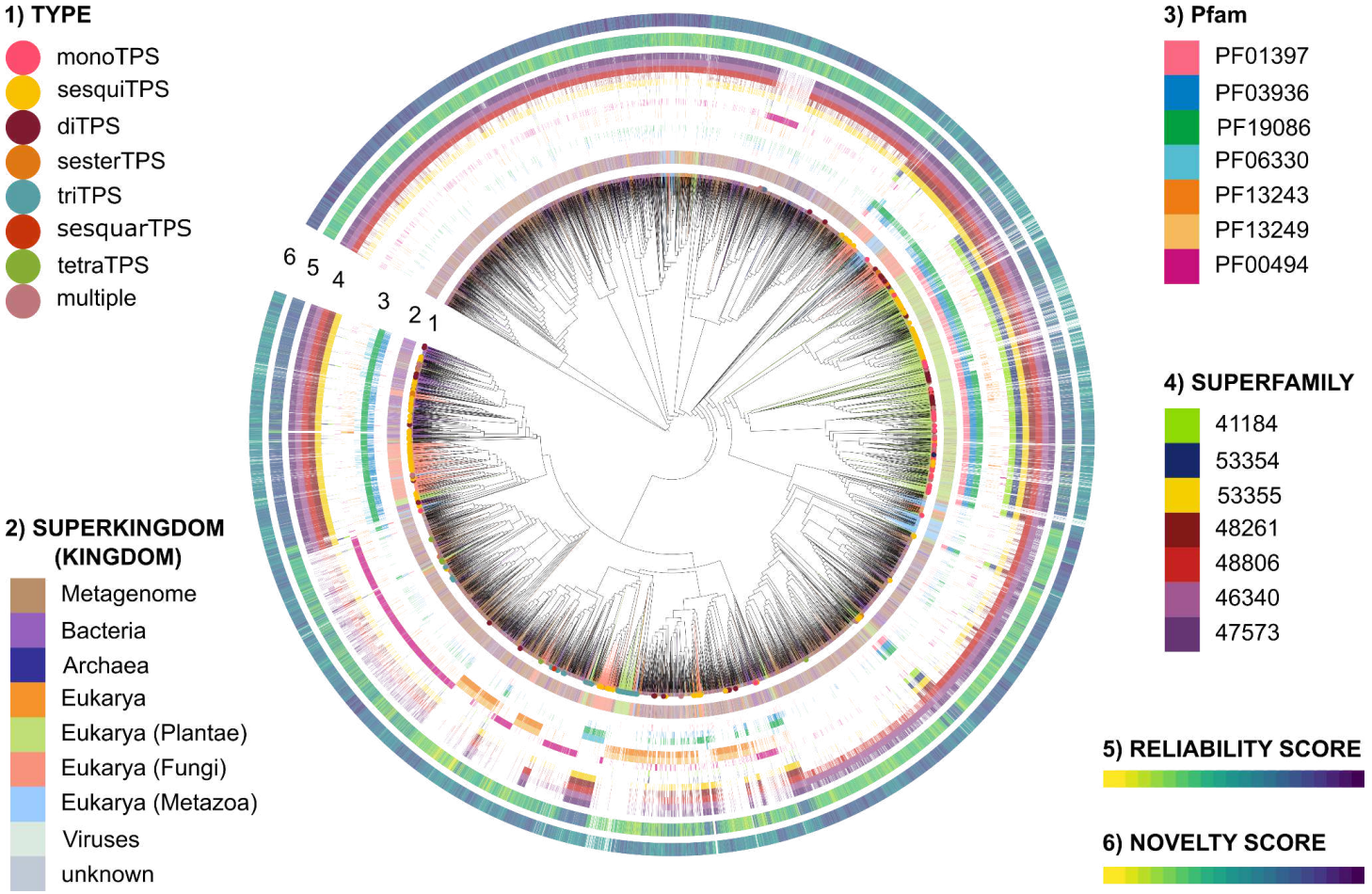

Figure 35. Phylogenetic tree of TPS candidates (representative sequences after 50% sequence identity clustering) and characterized TPSs. For clarity, the tree is displayed without branch lengths. Annotations around the phylogenetic tree include (1) TPS type for characterized TPSs, (2) superkingdom/kingdom, (3) Pfam domain presence, (4) SUPERFAMILY domain presence, (5) reliability score (0-6), and (6) novelty score (0-4).

5.3 Exploration of the TPS space using SSNs

To visualize the sequence space, the SSN from Chapter 4.6.2.2 is presented here, along with various annotations (see attachment A.4). The SSN nicely visualizes which parts of sequence space have not been explored so far (Figure 36).

Similar to the phylogenetic tree, there is a large cluster formed primarily by bacterial and metagenomic sequences and there are numerous smaller clusters without any characterized sequences. These parts of the TPS sequence space represent the potential for finding novel TPSs and terpenes.

The advantage of SSN visualization is that, firstly, it can be nicely seen how densely the clusters are connected within. Secondly, it can be nicely seen how the clusters are related.

From the characterized sequences in the network, we can see that they cluster both based on type (Figure 36) and kingdom/superkingdom (Figure A1). Clustering based on kingdom/superkingdom is also observable for the uncharacterized sequences (Figure A1).

Similarly to the phylogenetic tree, both novelty (Figure A2) and reliability (Figure A3) scores are also visualized. Generally, the clusters containing characterized sequences have higher reliability scores, whereas clusters without characterized sequences score higher with novelty scores.

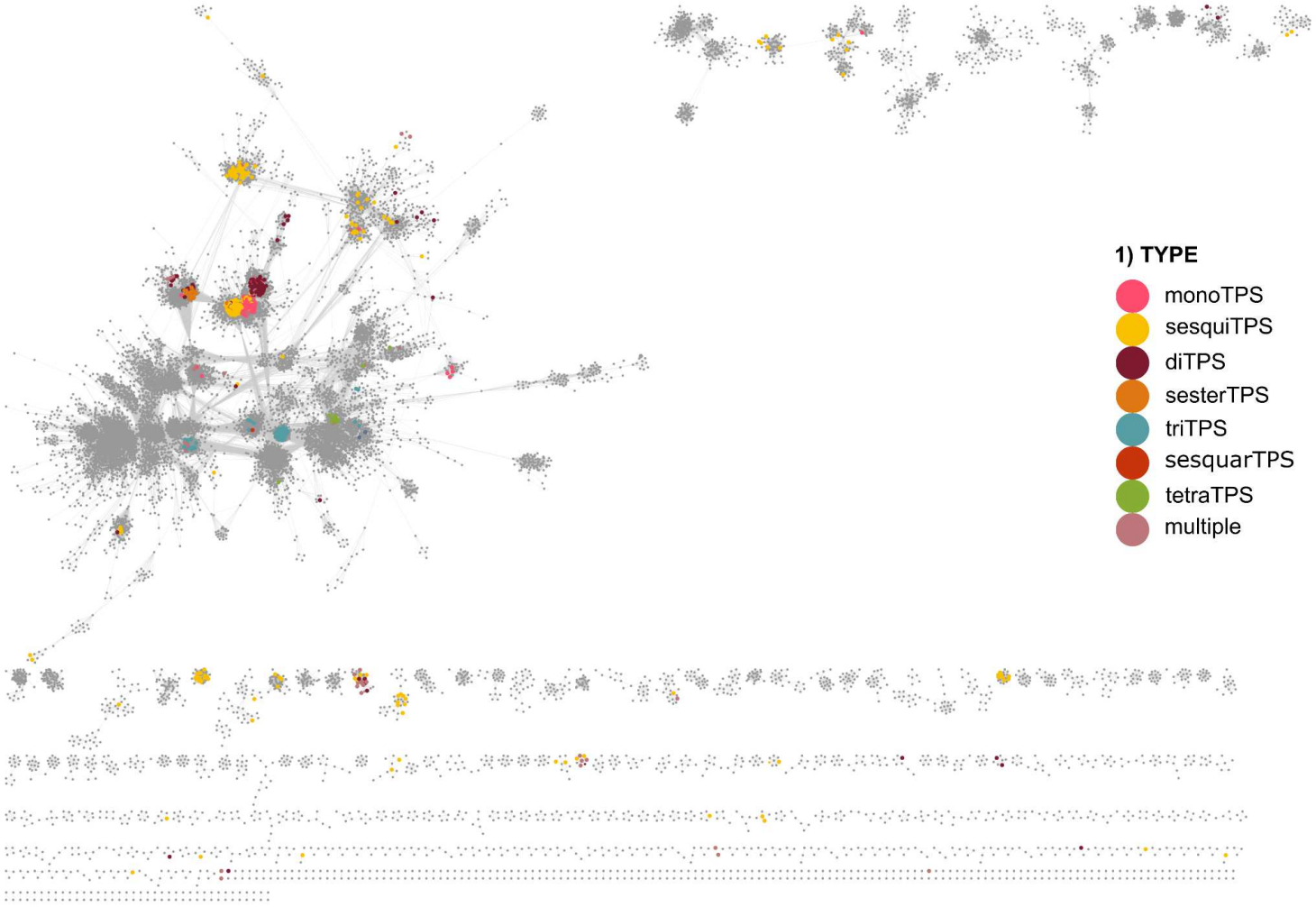

Figure 36. SSN with characterized TPSs colored according to their type, revealing distinct clustering patterns corresponding to different TPS types and uncovering unexplored regions of the TPS space.

6. Discussion

This thesis aimed to analyze characterized terpene synthases and utilize this knowledge to mine putative terpene synthases from large protein sequence repositories. Using Pfam domains for TPS mining is a standard method in the field, whereas using SUPERFAMILY domains for this purpose is less common. The employed filtering criteria, including sequence length, the presence of functional motifs, and domain completeness, align with established practices in the field.

The results were presented using two approaches: a phylogenetic tree and a sequence similarity network. Phylogenetic trees represent an established method in the field, while SSNs are a relatively newer method that has been gaining popularity in recent years.

The mining of terpene synthases revealed over 600 thousand sequences spanning various taxa. However, most sequences originate from metagenomes and bacteria. Notably, the exploration using the phylogenetic tree (Figure 35) and the SSN (Figure A1) revealed numerous large uncharacterized clades/clusters consisting mostly of bacterial and metagenomic sequences, highlighting their potential for further investigation as they also score high with the novelty score. Another interesting group of candidates is archaeal candidates, as Archaea can often be found in extreme environments (Wang et al. 2019), and there are currently no characterized archaeal TPSs in TPS db. In Eukaryotes, putative TPSs were newly identified, for example in *Bacillariophyta* (Diatoms), *Nematoda* (roundworms), or *Mollusca*. In addition, classes containing putative TPSs were newly identified, for example, in *Arthropoda*, *Chordata*, *Chlorophyta*, *Basidiomycota*, *Streptophyta*, or *Ascomycota*. More than 70 thousand TPS candidates were also identified in plants where the most interest could be directed using the novelty score, or it could be directed towards candidates from species or genera where TPSs have not been documented so far.

Although established methods were used, there are several limitations to the thesis that are further discussed. Firstly, in Chapter 4.1.3, the precision of the individual Pfam and SUPERFAMILY models was estimated, and models with low precision estimates were

discarded for further usage. Still, since the precision of the models was not 100%, false positives were likely incorporated. Subsequent filtering aimed to remove them. One of the filtering criteria was based on the presence of conserved functional motifs, which is a standard approach in the field. In Chapter 4.2.4, the characterized TPSs were analyzed for the presence of functional motifs, revealing that while most contain typical functional motifs, the majority of tri-, sesquar-, and tetraTPSs do not contain these motifs or may contain variations that were not detected. Therefore, filtering based on the presence of functional motifs could have potentially removed functional TPSs that do not contain these motifs or contain variations of them.

For mining of the BFD database, only clusters where the cluster representative contained some hit were inspected. However, this simplification may have potentially missed some sequences.

The majority of the mined results contain only SUPERFAMILY domain hits. It is interesting to see this big difference in the number of sequences captured by Pfam and SUPERFAMILY and the limited overlap of the hits. Characterized TPSs typically contain both Pfam and SUPERFAMILY domains, as shown in Chapter 4.2.1. Therefore, more reliable candidate sequences containg both Pfam and SUPERFAMILY domains. However, it would be interesting to experimentally test whether sequences containing only SUPERFAMILY domains exhibit TPS activity, as most of these sequences have multiple hits from the TPS SUPERFAMILY db.

Another limitation of using HMMs from Pfam and SUPERFAMILY is the possibility that completely novel TPSs exist beyond the detection capabilities of these models. For example, in insects, TPSs have evolved more recently, and their detection using Pfam models is not possible. In TPS db, five TPSs are not detectable by any Pfam or SUPERFAMILY model.

Compared to characterized TPSs, mined sequences tend to be shorter. Most of the sequences originate from bacteria and metagenomes, and hence likely from bacteria. This observation can be explained by the fact that bacteria generally produce shorter proteins, including TPSs, as shown in Chapter 4.2.2. The characterized TPS db mainly consists of plant TPSs having different lengths and domain architectures than TPSs

from other kingdoms and superkingdoms, such as bacteria. Another explanation could be that the mined sequences are fragmented.

Another potential source of caution is the quality of the sequence databases and their annotations. Metagenomic databases may contain lower-quality data (incomplete, fragmented etc.) compared to repositories like 1KP and Phytozome. The length filtering removed more than 70% of initially mined sequences from the metagenomic databases, indicating that many sequences in these databases are incomplete or fragmented. Another risk lies in the annotations, especially the taxonomical annotations, as some sequences could potentially originate from microbiomes or symbionts but were erroneously assigned as the host organism.

The exploration of the sequence space employed two methods: phylogenetic tree and SSN. Both methods have advantages and limitations, as previously discussed in chapters 2.2.1 and 2.2.2.

The construction of phylogenetic trees requires considerable expertise and poses many challenges, especially for large and diverse datasets like the one presented here. The primary objective of this thesis was not to conduct a comprehensive phylogenetic analysis. Instead, the phylogenetic tree presented here serves primarily as a tool for visualizing the relationships within the dataset. Building an MSA for this dataset is complicated and prone to errors. The MSA was automatically trimmed, which could possibly remove important positions. As the aim was to use the phylogenetic tree primarily for visualization, bootstrapping for estimating the branching confidence was omitted.

For SSN, several other challenges and limitations exist. One challenge is selecting the threshold for connecting nodes. While the separation of characterized sequences based on their type can be observed, uncertainty remains regarding whether this also applies to uncharacterized sequences. Additionally, SSNs cannot capture conserved functional residues, and the sequence identity metric cannot account for functionally similar proteins with dissimilar sequences.

To address these limitations of sequence identity as a similarity metric, there was an effort to use protein embedding distances for SSN instead. However, there was not informative enough separation into clusters (data not shown), potentially due to the low dimensionality of the selected model's embeddings. Nonetheless, exploring more complex models may be a promising direction for future improvement.

Additionally, while reliability and novelty scores provide valuable insights, they should be viewed as complementary rather than definitive selection criteria. The proposed novelty score was evaluated on TPS db to assess the correlation between the novelty score and the count of the most unique product (see Figure 38). TPS db was split into train (0.8) and test (0.2) sets. The train set was used to set the novelty score subscores for the test set. This assessment of the test set revealed a weak association (Spearman's rank correlation -0.44), suggesting that although the novelty score offers some insight, its ability to predict interesting products is limited.

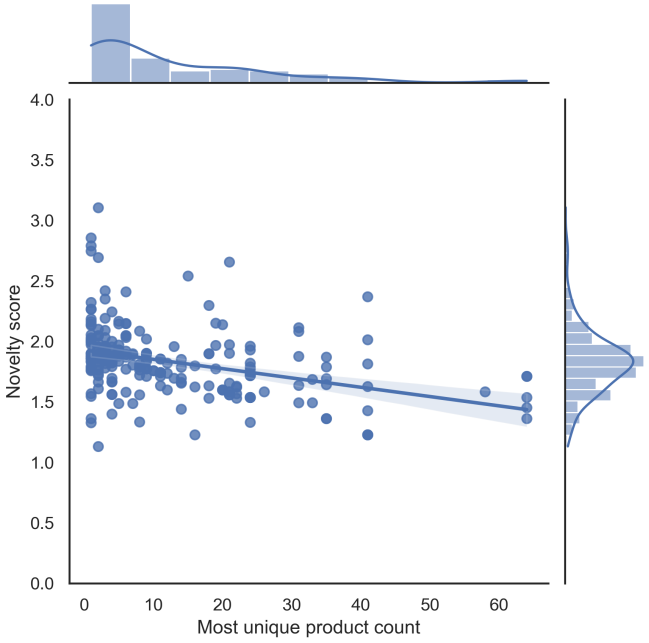

Figure 37. Comparison of novelty score and the count of the most unique product on the test set from TPS db.

Despite these limitations, the results presented here offer a rich source of putative terpene synthases for various applications. One straightforward application is the

exploration of the terpene synthase space for characterizing novel TPSs. Another application is in machine learning, as shown in (Samusevich et al. 2024), where this mined dataset was used to fine-tune a protein language model for TPS substrate prediction.

Prior to this thesis, the author conducted a similar TPS mining project. Nine candidates spanning plants, bacteria, and fungi were selected for experimental characterization detailed in (Smrčková 2023). Among the nine candidates, five exhibited terpene synthase activity, and from a fungal TPS, two sesquiterpenes are likely novel terpene compounds. This highlights the potential of the data gathered in this thesis to uncover additional novel terpene compounds.

Conclusions

Terpene synthases are fascinating enzymes responsible for the initial biosynthetic steps towards terpenoids, compounds with broad applications across various fields, including several applications in the pharmaceutical industry. The chemical synthesis of terpenoids presents a significant challenge, highlighting the importance of terpene synthases, which can be engineered into various host organisms to provide an alternative method for terpene production. This thesis primarily focused on two objectives: Firstly, to comprehensively analyze all experimentally characterized terpene synthases with bioinformatic methods, and subsequently, to leverage this knowledge to systematically mine large-scale sequence repositories for the discovery of novel terpene synthases.

The sequence-guided mining conducted in this thesis is, to my best knowledge, the largest reported effort to date, using nearly 5.5 billion sequences and identifying over 600 thousand putative novel terpene synthases spanning various taxa. The mining revealed that bacteria and metagenomes offer a rich reservoir of putative terpene synthases for further investigation. Furthermore, this mining led to the discovery of putative terpene synthases in taxa where they had not been previously documented. The resulting dataset serves as a valuable resource for the experimental characterization of novel terpene synthases and terpenes, as demonstrated in (Smrčková 2023). Additionally, this thesis facilitates efficient exploration of the putative terpene synthases by using the constructed phylogenetic tree, sequence similarity network, and two prioritization scores. Moreover, the dataset presented here can be utilized for various machine learning applications in the area of terpene biosynthesis, as demonstrated by the work of (Samusevich et al. 2024) on the problem of terpene synthase substrate prediction.

This bioinformatic analysis of all characterized terpene synthases has advanced our understanding of these enzymes. The revealed diversity of terpene synthases presents an exciting opportunity for further research and exploration, laying the groundwork for discovering novel terpenes.

References

Akiva, Eyal, Janine N. Copp, Nobuhiko Tokuriki, and Patricia C. Babbitt. 2017. "Evolutionary and Molecular Foundations of Multiple Contemporary Functions of the Nitroreductase Superfamily." *Proceedings of the National Academy of Sciences of the United States of America* 114 (45): E9549–58.

Alquézar, Berta, Ana Rodríguez, Marcos de la Peña, and Leandro Peña. 2017. "Genomic Analysis of Terpene Synthase Family and Functional Characterization of Seven Sesquiterpene Synthases from Citrus Sinensis." *Frontiers in Plant Science* 8 (August): 1481.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215 (3): 403–10.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs." *Nucleic Acids Research* 25 (17): 3389–3402.

Anaconda Inc. 2020. *Anaconda, Software Distribution*. https://anaconda.com/.

Apweiler, R., T. K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, et al. 2001. "The InterPro Database, an Integrated Documentation Resource for Protein Families, Domains and Functional Sites." *Nucleic Acids Research* 29 (1): 37–40.

Atallah, Christian, Katherine James, Zhen Ou, James Skelton, David Markham, Matt S. Burridge, James Finnigan, Simon Charnock, and Anil Wipat. 2024. "A Method for the Systematic Selection of Enzyme Panel Candidates by Solving the Maximum Diversity Problem." *Bio Systems* 236 (February): 105105.

Atkinson, Holly J., John H. Morris, Thomas E. Ferrin, and Patricia C. Babbitt. 2009. "Using Sequence Similarity Networks for Visualization of Relationships across Diverse Protein Superfamilies." *PloS One* 4 (2): e4345.

Aubourg, S., A. Lecharny, and J. Bohlmann. 2002. "Genomic Analysis of the Terpenoid Synthase (AtTPS) Gene Family of Arabidopsis Thaliana." *Molecular Genetics and Genomics: MGG* 267 (6): 730–45.

Boutanaev, Alexander M., Tessa Moses, Jiachen Zi, David R. Nelson, Sam T. Mugford, Reuben J. Peters, and Anne Osbourn. 2015. "Investigation of Terpene Diversification across Multiple Sequenced Plant Genomes." *Proceedings of the National Academy of Sciences of the United States of America* 112 (1): E81–88.

Brown, Duncan, and Kimmen Sjölander. 2006. "Functional Classification Using Phylogenomic Inference." *PLoS Computational Biology* 2 (6): e77.

Burkhardt, Immo, Tristan de Rond, Percival Yang-Ting Chen, and Bradley S. Moore. 2022. "Ancient Plant-like Terpene Biosynthesis in Corals." *Nature Chemical Biology* 18 (6): 664–69.

Cane, David E., and Haruo Ikeda. 2012. "Exploration and Mining of the Bacterial Terpenome." *Accounts of Chemical Research* 45 (3): 463–72.

Capella-Gutiérrez, Salvador, José M. Silla-Martínez, and Toni Gabaldón. 2009. "trimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses." *Bioinformatics* 25 (15): 1972–73.

Carpenter, Eric J., Naim Matasci, Saravanaraj Ayyampalayam, Shuangxiu Wu, Jing Sun, Jun Yu, Fabio Rocha Jimenez Vieira, et al. 2019. “Access to RNA-Sequencing Data from 1,173 Plant Species: The 1000 Plant Transcriptomes Initiative (1KP).” *GigaScience* 8 (10). https://doi.org/10.1093/gigascience/giz126.

Chen, Feng, Dorothea Tholl, Jörg Bohlmann, and Eran Pichersky. 2011. “The Family of Terpene Synthases in Plants: A Mid-Size Family of Genes for Specialized Metabolism That Is Highly Diversified throughout the Kingdom.” *The Plant Journal: For Cell and Molecular Biology* 66 (1): 212–29.

Chen, Rong, Qidong Jia, Xin Mu, Ben Hu, Xiang Sun, Zixin Deng, Feng Chen, Guangkai Bian, and Tiangang Liu. 2021. “Systematic Mining of Fungal Chimeric Terpene Synthases Using an Efficient Precursor-Providing Yeast Chassis.” *Proceedings of the National Academy of Sciences of the United States of America* 118 (29). https://doi.org/10.1073/pnas.2023247118.

Chen, Shanchong, Chi Zhang, and Lihan Zhang. 2022. “Investigation of the Molecular Landscape of Bacterial Aromatic Polyketides by Global Analysis of Type II Polyketide Synthases.” *Angewandte Chemie* 61 (24): e202202286.

Chen, Xinlu, Tobias G. Köllner, Qidong Jia, Ayla Norris, Balaji Santhanam, Patrick Rabe, Jeroen S. Dickschat, Gad Shaulsky, Jonathan Gershenzon, and Feng Chen. 2016. “Terpene Synthase Genes in Eukaryotes beyond Plants and Fungi: Occurrence in Social Amoebae.” *Proceedings of the National Academy of Sciences of the United States of America* 113 (43): 12132–37.

Chen, Zequn, Kelly J. Vining, Xiwu Qi, Xu Yu, Ying Zheng, Zhiqi Liu, Hailing Fang, et al. 2021. “Genome-Wide Analysis of Terpene Synthase Gene Family in Mentha Longifolia and Catalytic Activity Analysis of a Single Terpene Synthase.” *Genes* 12 (4). https://doi.org/10.3390/genes12040518.

Chhalodia, Anuj K., Houchao Xu, Georges B. Tabekoueng, Binbin Gu, Kizerbo A. Taizoumbe, Lukas Lauterbach, and Jeroen S. Dickschat. 2023. “Functional Characterisation of Twelve Terpene Synthases from Actinobacteria.” *Beilstein Journal of Organic Chemistry* 19 (September): 1386–98.

Christianson, David W. 2017. “Structural and Chemical Biology of Terpenoid Cyclases.” *Chemical Reviews* 117 (17): 11570–648.

Copp, Janine N., Eyal Akiva, Patricia C. Babbitt, and Nobuhiko Tokuriki. 2018. “Revealing Unexplored Sequence-Function Space Using Sequence Similarity Networks.” *Biochemistry* 57 (31): 4651–62.

Copp, Janine N., Dave W. Anderson, Eyal Akiva, Patricia C. Babbitt, and Nobuhiko Tokuriki. 2019. “Chapter Twelve - Exploring the Sequence, Function, and Evolutionary Space of Protein Superfamilies Using Sequence Similarity Networks and Phylogenetic Reconstructions.” In *Methods in Enzymology*, edited by Bruce A. Palfey, 620:315–47. Academic Press.

Eddy, Sean R. 2023. “HMMER.” Hmmer.org. August 2023. http://hmmer.org/

Eddy, S. R. 1998. “Profile Hidden Markov Models.” *Bioinformatics* 14 (9): 755–63.

Elnaggar, Ahmed, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, et al. 2022. “ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44 (10): 7112–27.

Finn, Robert D., Jaina Mistry, Benjamin Schuster-Böckler, Sam Griffiths-Jones, Volker Hollich, Timo Lassmann, Simon Moxon, et al. 2006. “Pfam: Clans, Web Tools and Services.” *Nucleic Acids Research* 34 (Database issue): D247–51.

Fu, Limin, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. 2012. “CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data.” *Bioinformatics* 28 (23): 3150–52.

Gao, Yang, Richard B. Honzatko, and Reuben J. Peters. 2012. “Terpenoid Synthase Structures: A so Far Incomplete View of Complex Catalysis.” *Natural Product Reports* 29 (10): 1153–75.

González-Hernández, Ricardo A., Norma A. Valdez-Cruz, Martha L. Macías-Rubalcava, and Mauricio A. Trujillo-Roldán. 2023. “Overview of Fungal Terpene Synthases and Their Regulation.” *World Journal of Microbiology & Biotechnology* 39 (7): 194.

Goodstein, David M., Shengqiang Shu, Russell Howson, Rochak Neupane, Richard D. Hayes, Joni Fazo, Therese Mitros, et al. 2012. “Phytozome: A Comparative Platform for Green Plant Genomics.” *Nucleic Acids Research* 40 (Database issue): D1178–86.

Gough, J., K. Karplus, R. Hughey, and C. Chothia. 2001. “Assignment of Homology to Genome Sequences Using a Library of Hidden Markov Models That Represent All Proteins of Known Structure.” *Journal of Molecular Biology* 313 (4): 903–19.

Haas, Brian J. 2022. *TransDecoder* (version 5.5.0). Github. https://github.com/TransDecoder/TransDecoder.

Hagberg, Aric, Pieter J. Swart, and Daniel A. Schult. 2008. “Exploring Network Structure, Dynamics, and Function Using NetworkX.” LA-UR-08-05495; LA-UR-08-5495. Los Alamos National Laboratory (LANL), Los Alamos, NM (United States). https://www.osti.gov/servlets/purl/960616.

Hage, Hayat, Julie Couillaud, Asaf Salamov, Margot Loussouarn-Yvon, Fabien Durbesson, Elena Ormeño, Sacha Grisel, et al. 2023. “An HMM Approach Expands the Landscape of Sesquiterpene Cyclases across the Kingdom Fungi.” *Microbial Genomics* 9 (4). https://doi.org/10.1099/mgen.0.000990.

Helfrich, Eric J. N., Geng-Min Lin, Christopher A. Voigt, and Jon Clardy. 2019. “Bacterial Terpene Biosynthesis: Challenges and Opportunities for Pathway Engineering.” *Beilstein Journal of Organic Chemistry* 15 (November): 2889–2906.

Hon, Jiri, Simeon Borko, Jan Stourac, Zbynek Prokop, Jaroslav Zendulka, David Bednar, Tomas Martinek, and Jiri Damborsky. 2020. “EnzymeMiner: Automated Mining of Soluble Enzymes with Diverse Structures, Catalytic Properties and Stabilities.” *Nucleic Acids Research* 48 (W1): W104–9.

Huang, Zheng-Yu, Ru-Yi Ye, Hui-Lei Yu, Ai-Tao Li, and Jian-He Xu. 2021. “Mining Methods and Typical Structural Mechanisms of Terpene Cyclases.” *Bioresources and Bioprocessing* 8 (1): 1–27.

Hu, Yi Ling, Qi Zhang, Shuang He Liu, Jia Li Sun, Fang Zhou Yin, Zi Ru Wang, Jing Shi, Rui Hua Jiao, and Hui Ming Ge. 2023. “Building Streptomyces Albus as a Chassis for Synthesis of Bacterial Terpenoids.” *Chemical Science* 14 (13): 3661–67.

Jiang, Kaibin, Chengju Du, Linwang Huang, Jiexian Luo, Tianyi Liu, and Shaowei Huang. 2023. “Phylotranscriptomics and Evolution of Key Genes for Terpene Biosynthesis in Pinaceae.” *Frontiers in Plant Science* 14 (February): 1114579.

Jiang, Shu-Ye, Jingjing Jin, Rajani Sarojam, and Srinivasan Ramachandran. 2019. “A Comprehensive Survey on the Terpene Synthase Gene Family Provides New Insight into Its Evolutionary Patterns.” *Genome Biology and Evolution* 11 (8): 2078–98.

Jia, Qidong, Reid Brown, Tobias G. Köllner, Jianyu Fu, Xinlu Chen, Gane Ka-Shu Wong, Jonathan Gershenzon, Reuben J. Peters, and Feng Chen. 2022. "Origin and Early Evolution of the Plant Terpene Synthase Family." *Proceedings of the National Academy of Sciences of the United States of America* 119 (15): e2100361119.
Jia, Qidong, Tobias G. Köllner, Jonathan Gershenzon, and Feng Chen. 2018. "MTPSLs: New Terpene Synthases in Nonseed Plants." *Trends in Plant Science* 23 (2): 121–28.
Jia, Qidong, Guanglin Li, Tobias G. Köllner, Jianyu Fu, Xinlu Chen, Wangdan Xiong, Barbara J. Crandall-Stotler, et al. 2016. "Microbial-Type Terpene Synthase Genes Occur Widely in Nonseed Land Plants, but Not in Seed Plants." *Proceedings of the National Academy of Sciences of the United States of America* 113 (43): 12328–33.
Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. "Highly Accurate Protein Structure Prediction with AlphaFold." *Nature* 596 (7873): 583–89.
Jung, Youngcheol, Takaaki Mitsuhashi, Sota Sato, Miki Senda, Toshiya Senda, and Makoto Fujita. 2023. "Function and Structure of a Terpene Synthase Encoded in a Giant Virus Genome." *Journal of the American Chemical Society* 145 (48): 25966–70.
Kang, Hahk-Soo, and Sean F. Brady. 2014. "Mining Soil Metagenomes to Better Understand the Evolution of Natural Product Structural Diversity: Pentangular Polyphenols as a Case Study." *Journal of the American Chemical Society* 136 (52): 18111–19.
Köster, Johannes, and Sven Rahmann. 2012. "Snakemake--a Scalable Bioinformatics Workflow Engine." *Bioinformatics* 28 (19): 2520–22.
Krogh, A., M. Brown, I. S. Mian, K. Sjölander, and D. Haussler. 1994. "Hidden Markov Models in Computational Biology. Applications to Protein Modeling." *Journal of Molecular Biology* 235 (5): 1501–31.
Kryukov, Kirill. 2021. *Fasta-Splitter* (version 0.2.6). http://kirill-kryukov.com/study/tools/fasta-splitter/.
Külheim, Carsten, Amanda Padovan, Charles Hefer, Sandra T. Krause, Tobias G. Köllner, Alexander A. Myburg, Jörg Degenhardt, and William J. Foley. 2015. "The Eucalyptus Terpene Synthase Gene Family." *BMC Genomics* 16 (1): 450.
Leferink, Nicole G. H., and Nigel S. Scrutton. 2022. "Predictive Engineering of Class I Terpene Synthases Using Experimental and Computational Approaches." *Chembiochem: A European Journal of Chemical Biology* 23 (5): e202100484.
Letunic, Ivica, and Peer Bork. 2021. "Interactive Tree Of Life (iTOL) v5: An Online Tool for Phylogenetic Tree Display and Annotation." *Nucleic Acids Research* 49 (W1): W293–96.
Lima Morais, David A. de, Hai Fang, Owen J. L. Rackham, Derek Wilson, Ralph Pethica, Cyrus Chothia, and Julian Gough. 2011. "SUPERFAMILY 1.75 Including a Domain-Centric Gene Ontology Method." *Nucleic Acids Research* 39 (Database issue): D427–34.
Lin, Zeming, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, et al. 2023. "Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model." *Science* 379 (6637): 1123–30.
Liu, Yanbin, Xixian Chen, and Congqiang Zhang. 2023. "Sustainable Biosynthesis of Valuable Diterpenes in Microbes." *Engineering Microbiology* 3 (1): 100058.
Li, Weizhong, and Adam Godzik. 2006. "Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences." *Bioinformatics* 22 (13): 1658–59.

Li, Zining, Baofu Xu, Volga Kojasoy, Teresa Ortega, Donovon A. Adpressa, Wenbo Ning, Xiuting Wei, et al. 2023. “First Trans-Eunicellane Terpene Synthase in Bacteria.” *Chem* 9 (3): 698–708.

Maaten, L., and Geoffrey E. Hinton. 2008. “Visualizing Data Using T-SNE.” *Journal of Machine Learning Research: JMLR* 9: 2579–2605.

Madera, Martin, and Julian Gough. 2002. “A Comparison of Profile Hidden Markov Model Procedures for Remote Homology Detection.” *Nucleic Acids Research* 30 (19): 4321–28.

Malit, Jessie James Limlingan, Hiu Yu Cherie Leung, and Pei-Yuan Qian. 2022. “Targeted Large-Scale Genome Mining and Candidate Prioritization for Natural Product Discovery.” *Marine Drugs* 20 (6). https://doi.org/10.3390/md20060398.

Martin, Diane M., Sébastien Aubourg, Marina B. Schouwey, Laurent Daviet, Michel Schalk, Omid Toub, Steven T. Lund, and Jörg Bohlmann. 2010. “Functional Annotation, Genome Organization and Phylogeny of the Grapevine (Vitis Vinifera) Terpene Synthase Gene Family Based on Genome Assembly, FLcDNA Cloning, and Enzyme Assays.” *BMC Plant Biology* 10 (October): 226.

McInnes, Leland, John Healy, and James Melville. 2018. “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.” *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1802.03426.

Mistry, Jaina, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A. Salazar, Erik L. L. Sonnhammer, Silvio C. E. Tosatto, et al. 2021. “Pfam: The Protein Families Database in 2021.” *Nucleic Acids Research* 49 (D1): D412–19.

Moosmann, Philipp, Felix Ecker, Stefan Leopold-Messer, Jackson K. B. Cahn, Cora L. Dieterich, Michael Groll, and Jörn Piel. 2020. “A Monodomain Class II Terpene Cyclase Assembles Complex Isoprenoid Scaffolds.” *Nature Chemistry* 12 (10): 968–72.

Murzin, A. G., S. E. Brenner, T. Hubbard, and C. Chothia. 1995. “SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures.” *Journal of Molecular Biology* 247 (4): 536–40.

Oberg, Nils, Rémi Zallot, and John A. Gerlt. 2023. “EFI-EST, EFI-GNT, and EFI-CGFP: Enzyme Function Initiative (EFI) Web Resource for Genomic Enzymology Tools.” *Journal of Molecular Biology* 435 (14): 168018.

One Thousand Plant Transcriptomes Initiative. 2019. “One Thousand Plant Transcriptomes and the Phylogenomics of Green Plants.” *Nature* 574 (7780): 679–85.

Pandurangan, Arun Prasad, Jonathan Stahlhacke, Matt E. Oates, Ben Smithers, and Julian Gough. 2019. “The SUPERFAMILY 2.0 Database: A Significant Proteome Update and a New Webserver.” *Nucleic Acids Research* 47 (D1): D490–94.

Park, J., K. Karplus, C. Barrett, R. Hughey, D. Haussler, T. Hubbard, and C. Chothia. 1998. “Sequence Comparisons Using Multiple Sequences Detect Three Times as Many Remote Homologues as Pairwise Methods.” *Journal of Molecular Biology* 284 (4): 1201–10.

Paysan-Lafosse, Typhaine, Matthias Blum, Sara Chuguransky, Tiago Grego, Beatriz Lázaro Pinto, Gustavo A. Salazar, Maxwell L. Bileschi, et al. 2023. “InterPro in 2022.” *Nucleic Acids Research* 51 (D1): D418–27.

Price, Morgan N., Paramvir S. Dehal, and Adam P. Arkin. 2010. “FastTree 2--Approximately Maximum-Likelihood Trees for Large Alignments.” *PloS One* 5 (3): e9490.

Quin, Maureen B., Christopher M. Flynn, and Claudia Schmidt-Dannert. 2014. “Traversing the Fungal Terpenome.” *Natural Product Reports* 31 (10): 1449–73.

Rebholz, Zarley, Leena Shewade, Kylie Kaler, Hailey Larose, Florian Schubot, Dorothea Tholl, Alexandre V. Morozov, and Paul E. O'Maille. 2023. "Emergence of Terpene Chemical Communication in Insects: Evolutionary Recruitment of Isoprenoid Metabolism." *Protein Science: A Publication of the Protein Society* 32 (5): e4634.
Reddy, Gajendar Komati, Nicole G. H. Leferink, Maiko Umemura, Syed T. Ahmed, Rainer Breitling, Nigel S. Scrutton, and Eriko Takano. 2020. "Exploring Novel Bacterial Terpene Synthases." *PloS One* 15 (4): e0232220.
Rice, P., I. Longden, and A. Bleasby. 2000. "EMBOSS: The European Molecular Biology Open Software Suite." *Trends in Genetics: TIG* 16 (6): 276–77.
Richardson, Lorna, Ben Allen, Germana Baldi, Martin Beracochea, Maxwell L. Bileschi, Tony Burdett, Josephine Burgin, et al. 2023. "MGnify: The Microbiome Sequence Data Analysis Resource in 2023." *Nucleic Acids Research* 51 (D1): D753–59.
Rives, Alexander, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, et al. 2021. "Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences." *Proceedings of the National Academy of Sciences of the United States of America* 118 (15). https://doi.org/10.1073/pnas.2016239118.
Rudolf, Jeffrey D., and Chin-Yuan Chang. 2020. "Terpene Synthases in Disguise: Enzymology, Structure, and Opportunities of Non-Canonical Terpene Synthases." *Natural Product Reports* 37 (3): 425–63.
Samusevich, Raman, Téo Hebra, Roman Bushuiev, Anton Bushuiev, Ratthachat Chatpatanasiri, Jonáš Kulhánek, Tereza Čalounová, et al. 2024. "Discovery and Characterization of Terpene Synthases Powered by Machine Learning." *bioRxiv*. https://doi.org/10.1101/2024.01.29.577750.
Sayers, Eric W., Mark Cavanaugh, Karen Clark, Kim D. Pruitt, Stephen T. Sherry, Linda Yankie, and Ilene Karsch-Mizrachi. 2023. "GenBank 2023 Update." *Nucleic Acids Research* 51 (D1): D141–44.
Scesa, Paul D., Zhenjian Lin, and Eric W. Schmidt. 2022. "Ancient Defensive Terpene Biosynthetic Gene Clusters in the Soft Corals." *Nature Chemical Biology* 18 (6): 659–63.
Schmidt-Dannert, Claudia. 2015. "Biosynthesis of Terpenoid Natural Products in Fungi." In *Biotechnology of Isoprenoids*, edited by Jens Schrader and Jörg Bohlmann, 19–61. Cham: Springer International Publishing.
Schwab, Wilfried, Christopher Fuchs, and Fong-Chin Huang. 2013. "Transformation of Terpenes into Fine Chemicals." *European Journal of Lipid Science and Technology: EJLST* 115 (1): 3–8.
Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks." *Genome Research* 13 (11): 2498–2504.
Sievers, Fabian, Andreas Wilm, David Dineen, Toby J. Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, et al. 2011. "Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega." *Molecular Systems Biology* 7 (October): 539.
Smrčková, Helena. 2023. "Terpene Discovery Combining in Silico and Molecular Biology Approaches." Edited by Pluskal Tomáš. Bc, Univerzita Karlova. http://hdl.handle.net/20.500.11956/181562.

Sonnhammer, E. L., S. R. Eddy, and R. Durbin. 1997. “Pfam: A Comprehensive Database of Protein Domain Families Based on Seed Alignments.” *Proteins* 28 (3): 405–20.
Sun, Yang, Wenqing Xiao, Qing-Nan Wang, Jing Wang, Xiang-Dong Kong, Wen-Hui Ma, Si-Xian Liu, Ping Ren, Li-Na Xu, and Yong-Jun Zhang. 2023. “Multiple Variation Patterns of Terpene Synthases in 26 Maize Genomes.” *BMC Genomics* 24 (1): 46.
“Taxoniq: Taxon Information Query - Fast, Offline Querying of NCBI Taxonomy and Related Data.” n.d. Accessed April 6, 2024. https://taxoniq.github.io/taxoniq/.
Tholl, Dorothea, Zarley Rebholz, Alexandre V. Morozov, and Paul E. O’Maille. 2023. “Terpene Synthases and Pathways in Animals: Enzymology and Structural Evolution in the Biosynthesis of Volatile Infochemicals.” *Natural Product Reports* 40 (4): 766–93.
UniProt Consortium. 2023. “UniProt: The Universal Protein Knowledgebase in 2023.” *Nucleic Acids Research* 51 (D1): D523–31.
Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. “Attention Is All You Need.” *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1706.03762.
Wang, Shengqin, Zhihong Zheng, Huixi Zou, Nan Li, and Mingjiang Wu. 2019. “Characterization of the Secondary Metabolite Biosynthetic Gene Clusters in Archaea.” *Computational Biology and Chemistry* 78 (February): 165–69.
Wei, Guo, Qidong Jia, Xinlu Chen, Tobias G. Köllner, Debashish Bhattacharya, Gane Ka-Shu Wong, Jonathan Gershenzon, and Feng Chen. 2019. “Terpene Biosynthesis in Red Algae Is Catalyzed by Microbial Type But Not Typical Plant Terpene Synthases.” *Plant Physiology* 179 (2): 382–90.
Wilson, Kayla, Tristan de Rond, Immo Burkhardt, Taylor S. Steele, Rebecca J. B. Schäfer, Sheila Podell, Eric E. Allen, and Bradley S. Moore. 2023. “Terpene Biosynthesis in Marine Sponge Animals.” *Proceedings of the National Academy of Sciences of the United States of America* 120 (9): e2220934120.
Yamada, Yuuki, Tomohisa Kuzuyama, Mamoru Komatsu, Kazuo Shin-Ya, Satoshi Omura, David E. Cane, and Haruo Ikeda. 2015. “Terpene Synthases Are Widely Distributed in Bacteria.” *Proceedings of the National Academy of Sciences of the United States of America* 112 (3): 857–62.
Yan, Xue-Mei, Shan-Shan Zhou, Hui Liu, Shi-Wei Zhao, Xue-Chan Tian, Tian-Le Shi, Yu-Tao Bao, et al. 2023. “Unraveling the Evolutionary Dynamics of the TPS Gene Family in Land Plants.” *Frontiers in Plant Science* 14 (October): 1273648.
Zallot, Rémi, Nils Oberg, and John A. Gerlt. 2019. “The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways.” *Biochemistry* 58 (41): 4169–82.
Zallot, Remi, Nils Oberg, and John A. Gerlt. 2021. “Discovery of New Enzymatic Functions and Metabolic Pathways Using Genomic Enzymology Web Tools.” *Current Opinion in Biotechnology* 69 (June): 77–90.
Zerbe, Philipp, and Jörg Bohlmann. 2015. “Plant Diterpene Synthases: Exploring Modularity and Metabolic Diversity for Bioengineering.” *Trends in Biotechnology* 33 (7): 419–28.
Zhang, Congqiang, Xixian Chen, Axel Orban, Sudha Shukal, Florian Birk, Heng-Phon Too, and Martin Rühl. 2020. “Agrocybe Aegerita Serves As a Gateway for Identifying Sesquiterpene Biosynthetic Enzymes in Higher Fungi.” *ACS Chemical Biology* 15 (5): 1268–77.

Zhao, Le, Yunhao Zhu, Haoyu Jia, Yongguang Han, Xiaoke Zheng, Min Wang, and Weisheng Feng. 2022. “From Plant to Yeast-Advances in Biosynthesis of Artemisinin.” *Molecules* 27 (20). https://doi.org/10.3390/molecules27206888.

List of figures

- **Figure 36.** SSN with characterized TPSs colored according to their type, revealing distinct clustering patterns corresponding to different TPS types and uncovering unexplored regions of the TPS space.
- **Figure 37.** Comparison of novelty score and the count of the most unique product on the test set from TPS db.
- **Figure A1.** SSN with sequences colored according to superkingdom/kingdom. Characterized TPSs are outlined in black.
- **Figure A2.** SSN with candidate sequences according to their novelty score (0=yellow, 4=purple). Characterized TPSs are colored according to their type.
- **Figure A3.** SSN with candidate sequences colored according to their reliability score (0=yellow, 6=dark purple). Characterized TPSs are colored according to their type.

List of tables

A. Attachments

A.1 SSN Figures

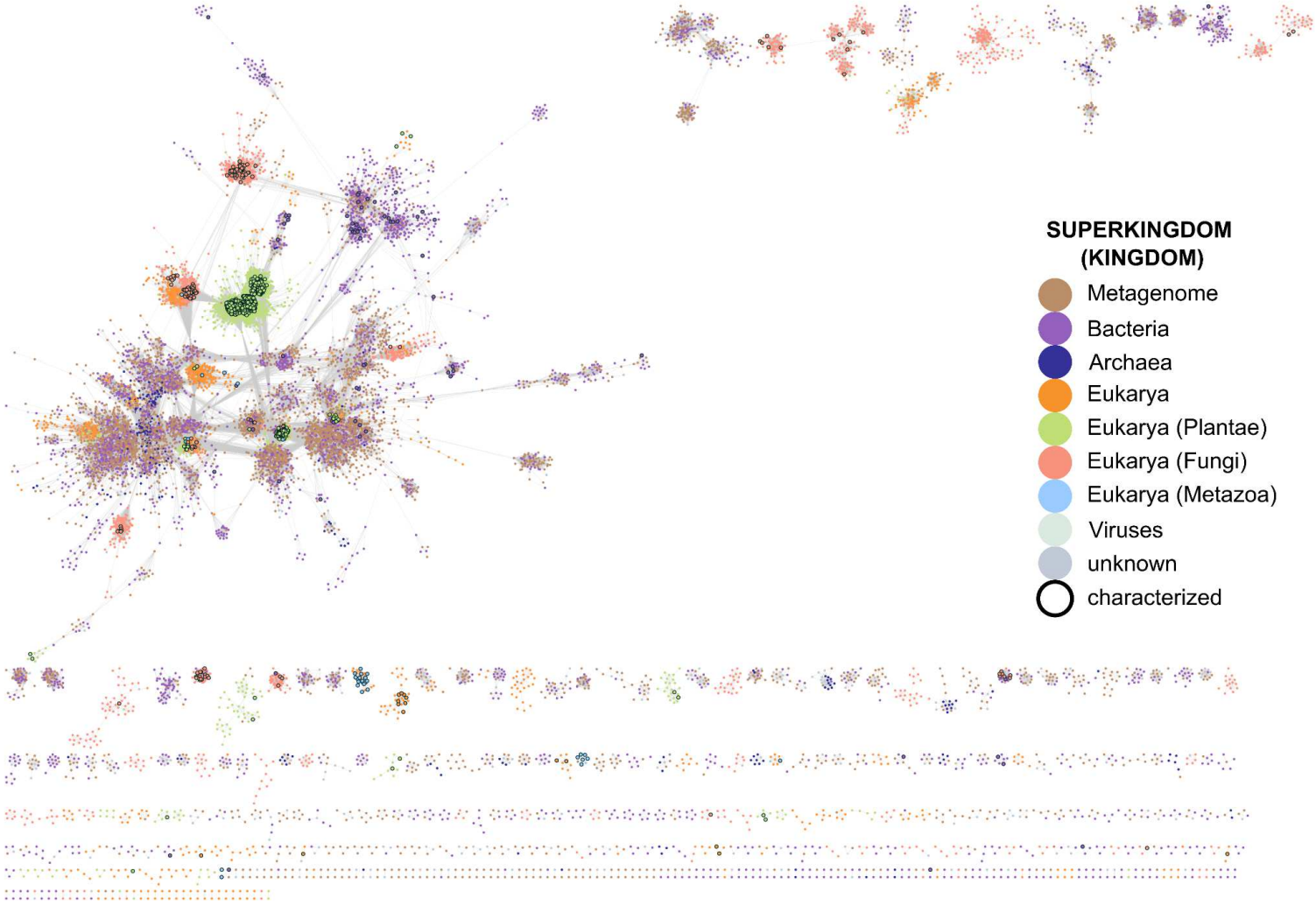

Figure A1. SSN with sequences colored according to superkingdom/kingdom. Characterized TPSs are outlined in black.

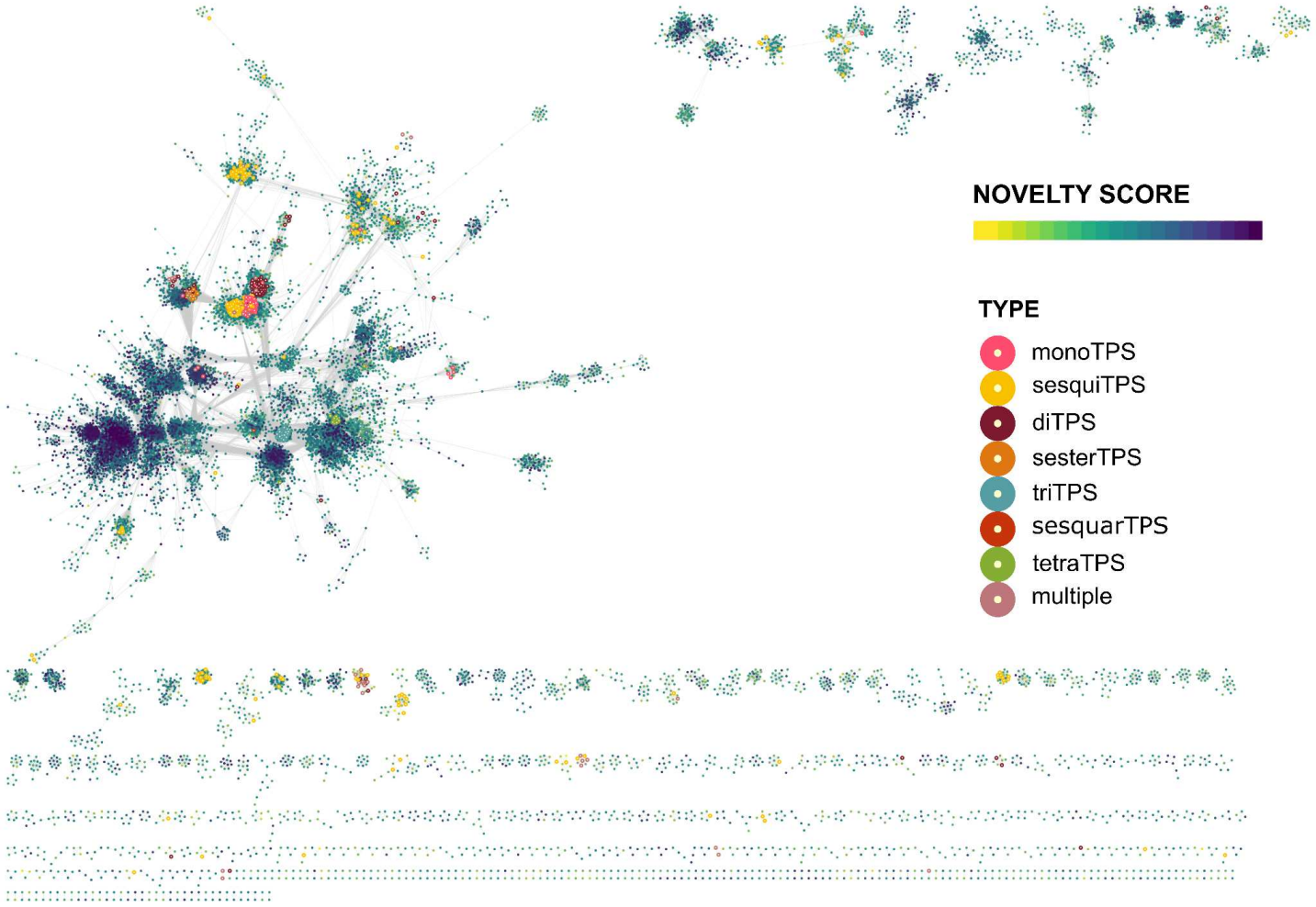

Figure A2. SSN with candidate sequences according to their novelty score (0=yellow, 4=purple). Characterized TPSs are colored according to their type.

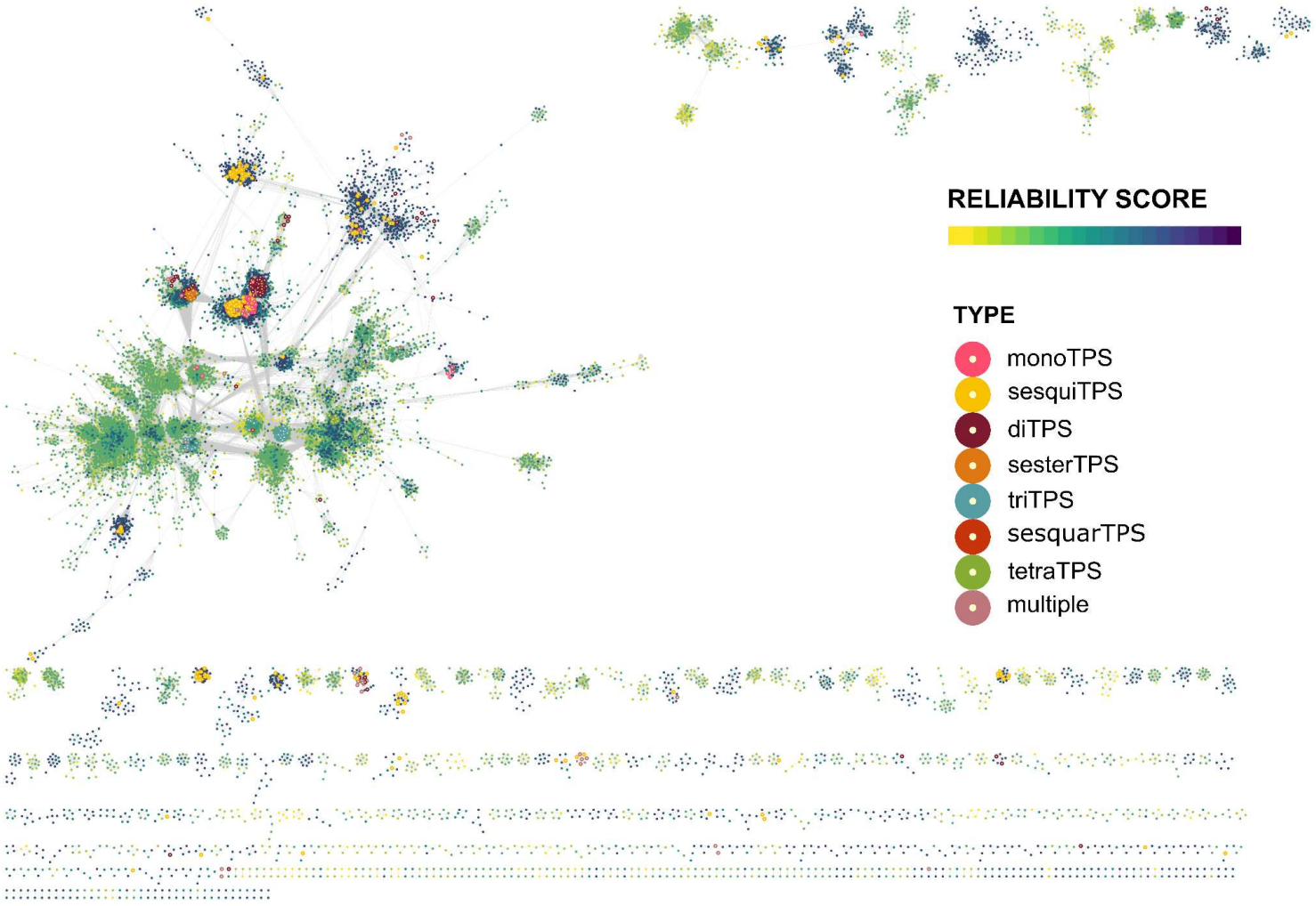

Figure A3. SSN with candidate sequences colored according to their reliability score (0=yellow, 6=dark purple). Characterized TPSs are colored according to their type.

A.2 TPS db dataset

The original TPS db dataset curated by the members of Pluskal lab at IOCB Prague is available as an annotated csv file named *tps_db_original.csv*. The dataset contains 2515 entries of individual reactions corresponding to 1323 proteins, including some IDSs.

This dataset was filtered and further annotated, as described in Chapter 4.2. This resulting dataset of characterized TPSs used for the analysis is available as an annotated csv file named *tps_db.csv*. The dataset contains 1025 entries corresponding to individual TPSs and the following attributes as columns:

- **id:** UniProt ID or other identifier
- **type:** type(s) (mono/sesq/di/...)
- **name:** UniProt name
- **sequence:** amino acid sequence
- **species:** species
- **kingdom:** kingdom
- **pfam_architecture:** Pfam architecture as a list of consecutive domains; domains covered by less than 50% have suffix "_partial"
- **supfam_architecture:** SUPERFAMILY architecture as a list of consecutive domains; domains covered by less than 50% have suffix "_partial"
- **PF06330.14:** a binary indicator (1 or 0) for the presence of Pfam domain PF06330.14
- **PF01397.24:** a binary indicator (1 or 0) for the presence of Pfam domain PF01397.24
- **PF03936.19:** a binary indicator (1 or 0) for the presence of Pfam domain PF03936.19
- **PF00494.22:** a binary indicator (1 or 0) for the presence of Pfam domain PF00494.22
- **PF13249.9:** a binary indicator (1 or 0) for the presence of Pfam domain PF13249.9
- **PF19086.3:** a binary indicator (1 or 0) for the presence of Pfam domain PF19086.3
- **PF13243.9:** a binary indicator (1 or 0) for the presence of Pfam domain PF13243.9

- **0041184:** a binary indicator (1 or 0) for the presence of SUPERFAMILY domain 0041184
- **0053354**: a binary indicator (1 or 0) for the presence of SUPERFAMILY domain 0053354
- **0053355**: a binary indicator (1 or 0) for the presence of SUPERFAMILY domain 0053355
- **0048261**: a binary indicator (1 or 0) for the presence of SUPERFAMILY domain 0048261
- **0048806**: a binary indicator (1 or 0) for the presence of SUPERFAMILY domain 0048806
- **0046340**: a binary indicator (1 or 0) for the presence of SUPERFAMILY domain 0046340
- **0047573**: a binary indicator (1 or 0) for the presence of SUPERFAMILY domain 0047573
- **motifs**: a list of present conserved motifs
- **DDXXD**: a binary indicator (True or False) for the presence of the DDXXD motif
- **NSE/DTE:** a binary indicator (True or False) for the presence of the NSE/DTE motif
- **DXDD:** a binary indicator (True or False) for the presence of the DXDD motif
- **length:** length of the amino acid sequence
- **tax_id:** NCBI taxon ID
- **lineage:** taxonomic lineage based on NCBI taxon ID
- **ncbi_species:** species based on NCBI taxon ID
- **ncbi_genus:** genus based on NCBI taxon ID
- **ncbi_family:** family based on NCBI taxon ID
- **ncbi_order:** order based on NCBI taxon ID
- **ncbi_class:** class based on NCBI taxon ID
- **ncbi_phylum:** phylum based on NCBI taxon ID
- **ncbi_kingdom:** kingdom based on NCBI taxon ID
- **ncbi_superkingdom:** superkingdom based on NCBI taxon ID

A.3 TPS candidates dataset

The final dataset of TPS candidates is available as an annotated csv file named *tps_mining_dataset.csv*.
The dataset contains 606,791 entries and the following attributes as columns:

- **id:** id of the sequence assigned during the mining, prefix corresponds to the source database
- **record_id:** original id(s) of the sequence
- **record_description:** the original description of the sequence when available
- **length:** length of the amino acid sequence
- **architecture_pfam:** Pfam architecture as a list of consecutive domains; domains covered by less than 50% have suffix "_partial"
- **PF06330.14:** a binary indicator (1 or 0) for the presence of Pfam domain PF06330.14
- **PF01397.24:** a binary indicator (1 or 0) for the presence of Pfam domain PF01397.24
- **PF03936.19:** a binary indicator (1 or 0) for the presence of Pfam domain PF03936.19
- **PF00494.22:** a binary indicator (1 or 0) for the presence of Pfam domain PF00494.22
- **PF13249.9:** a binary indicator (1 or 0) for the presence of Pfam domain PF13249.9
- **PF19086.3:** a binary indicator (1 or 0) for the presence of Pfam domain PF19086.3
- **PF13243.9:** a binary indicator (1 or 0) for the presence of Pfam domain PF13243.9
- **architecture_supfam**: SUPERFAMILY architecture as a list of consecutive domains; domains covered by less than 50% have suffix "_partial"
- **0041184:** a binary indicator (1 or 0) for the presence of SUPERFAMILY domain 0041184
- **0053354: a** binary indicator (1 or 0) for the presence of SUPERFAMILY domain
- **0053355: a** binary indicator (1 or 0) for the presence of SUPERFAMILY domain
- **0048261: a** binary indicator (1 or 0) for the presence of SUPERFAMILY domain
- **0048806: a** binary indicator (1 or 0) for the presence of SUPERFAMILY domain

- **0046340:** a binary indicator (1 or 0) for the presence of SUPERFAMILY domain
- **0047573: a** binary indicator (1 or 0) for the presence of SUPERFAMILY domain
- **organism:** source organism
- **tax_id:** NCBI taxon ID
- **lineage:** taxonomic lineage based on NCBI taxon ID
- **superkingdom:** superkingdom/metagenome
- **kingdom:** kingdom
- **phylum:** phylum
- **methionin:** a binary indicator (1 or 0) for the presence of starting methionine
- **observed_pfam_architecture:** a binary indicator (1 or 0) if the Pfam architecture was observed in TPS db
- **observed_supfam_architecture:** a binary indicator (1 or 0) if the SUPERFAMILY architecture was observed in TPS db
- **strongest_pfam_hit_cevalue:** smallest c-evalue (from HMMER) of all TPS Pfam db domain hits for the sequence
- **strongest_pfam_hit_cevalue_neg_log_norm:** normalized logarithm of strongest_pfam_hit_cevalue
- **strongest_supfam_hit_cevalue:** smallest c-evalue (from HMMER) of all TPS SUPERFAMILY db domain hits for the sequence
- **strongest_supfam_hit_cevalue_neg_log_norm:** normalized logarithm of strongest_supfam_hit_cevalue
- **has_pfam_hit:** a binary indicator (1 or 0) for the presence of any domain from TPS Pfam db
- **has_supfam_hit:** a binary indicator (1 or 0) for the presence of any domain from TPS SUPERFAMILY db
- **has_pfam_and_supfam_hit:** a binary indicator (1 or 0) for the presence of any domain from both TPS Pfam db and TPS SUPERFAMILY db
- **reliability_score:** reliability score with range from 0 to 6
- **tax_score:** taxonomic score with a range from 0 to 1
- **phylo_score:** phylogenetic score with a range from 0 to 1
- **ssn_score:** SSN score with a range from 0 to 1
- **product_score:** product score with a range from 0 to 1

- **novelty_score:** novelty score with a range from 0 to 4
- **closest_char_tps_sid:** ID of closest characterized TPS from TPS db by sequence identity
- **sid:** sequence identity corresponding to closest_char_tps_sid
- **closest_char_tps_esm_eucl_dist:** ID of closest characterized TPS from TPS db by Euclidean distance of their ESM2 embeddings
- **closest_char_tps_esm:** Euclidean distance corresponding to closest_char_tps_esm_eucl_dist
- **cluster:** cluster id of cluster from 50% sequence identity clustering
- **sequence:** amino acid sequence

A.4 Other

A GitHub repository for this project is available at:
https://github.com/CalounovaT/TPS_mining

This repository contains a *README.md* file with an overview of the repository and a *packages.txt* file, which contains a list of used conda and pip packages and their versions. The repository is organized into three subdirectories corresponding to different parts of this thesis:

- *01_tps_db_analysis* (Chapter 4.2): https://github.com/CalounovaT/TPS_mining/tree/main/01_tps_db_analysis
- *02_mining* (Chapter 4.3 and Chapter 4.4): https://github.com/CalounovaT/TPS_mining/tree/main/02_mining
- *03_mining_analysis* (Chapter 4.5, Chapter 4.6, and Chapter 5): https://github.com/CalounovaT/TPS_mining/tree/main/03_mining_analysis

Each subdirectory contains a *README.md* file with a description, a *Snakefile* file, which was used to process and generate data, and other additional scripts and files. Directories *01_tps_db_analysis* and *03_mining_analysis* also contain a subdirectory *notebooks* with various Python and R notebooks used to analyze data and create plots.

Lastly, the phylogenetic tree from chapters 4.6.2.3 and 5.2 is available in the directory *03_mining_analysis/phylogenetic_tree*, along with instructions for its visualization. https://github.com/CalounovaT/TPS_mining/tree/main/03_mining_analysis/phylogenetic_tree

Similarly, the SSN from chapters 4.6.2.2 and 5.3 is available in the directory *03_mining_analysis/SSN*, along with Cytoscape session files where the SSN is annotated and colored.
https://github.com/CalounovaT/TPS_mining/tree/main/03_mining_analysis/SSN)